DOCUMENT RESUME

ED 385 566 TM 024 003

AUTHOR Wingersky, Marilyn S.
TITLE Significant Improvements to LOGIST.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-22
PUB DATE Apr 92
NOTE 83p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS Ability; *Computer Software Development; *Estimation (Mathematics); *Item Response Theory; *Maximum Likelihood Statistics
IDENTIFIERS *Item Parameters; *LOGIST Computer Program; Three Parameter Model

ABSTRACT

The computer program LOGIST (Wingersky, Patrick, and Lord, 1988) estimates the item parameters and the examinee's abilities for Birnbaum's three-parameter logistic item response theory model using Newton's method for solving the joint maximum likelihood equations. In 1989, Martha Stocking discovered a problem with this procedure in that when the true item discriminations were used as starting values for the iteration procedure, item parameters were different from when the default starting value of one was used for the item discriminations. When a straight run to convergence was performed, the different initial starting values converged to the same item parameter estimates. This study investigated several methods for improving the automatic procedure, but when they failed to yield the necessary improvement, a method was devised that gives estimates nearly as good as those obtained from running to convergence. The method involves adding a step to get better initial parameter estimates for the automatic procedure. Abilities are grouped coarsely, and the grouped abilities are estimated iteratively, alternating between items and abilities until the maximum difference between the estimated item characteristic curves is less than some criterion. A new version of LOGIST, LOGIST7, has been produced. Four tables and 33 figures present analysis details. (Contains nine references.) (SLD)

RESEARCH REPORT

SIGNIFICANT IMPROVEMENTS TO *LOGIST*

Marilyn S. Wingersky

Educational Testing Service
Princeton, New Jersey
April 1992

Significant Improvements to LOGIST*

Marilyn S. Wingersky

January 1992

* This work was supported by ETS through the Program Research Planning Council.

Abstract

Stocking (1989) discovered a problem with the LOGIST estimation procedure. The automatic procedure in this program produced different item parameter estimates when the true item discriminations were used as starting values for the iteration procedure than when the default starting value of one for the item discriminations was used. When a straight run to convergence was performed, the different initial starting values converged to the same item parameter estimates.

This study investigated several methods for improving the automatic procedure. When these methods failed to give the improvement necessary, several methods were tried to obtain the same results as the run to convergence in a shorter amount of computer time. A method was devised that takes much less time and gives nearly as good, and in some cases better, estimates as estimates obtained from running to convergence. This method involves adding a step to get better initial item parameter estimates for the automatic procedure. In this step, the abilities are grouped very coarsely and the item parameters and the grouped abilities are estimated iteratively, alternating back and forth between items and abilities until the maximum difference between the estimated item characteristic curves is less than some criterion.

The new procedure gives item parameter estimates that are closer to the true values than the current 4-step method does. However, there is a definite nonlinear relationship between the estimated item parameters for the two methods after the parameters have been linearly transformed to the same scale. Consequently, in an ongoing series of calibrations, switching to this

procedure from the old procedure will produce a discontinuity in the parameter estimates in the same manner as would be caused by switching from LOGIST to BILOG.

The effects of putting a beta prior on c were also investigated. The results were not conclusive. An option to put a beta prior on c was added to the program.

This new method has been incorporated into a new version of LOGIST called LOGIST7.

Introduction

The computer program, LOGIST (Wingersky, Patrick and Lord, 1988) estimates the item parameters and the examinee abilities for Birnbaum's 3-parameter logistic item response theory model using Newton's method for solving the joint maximum likelihood equations. This is not an easy estimation problem. Newton's method requires some initial starting values for the parameters and iteratively solves for corrections to these values to obtain the solution to the joint likelihood equations. In this problem there are N+3n unknowns, where N is the number of examinees and n is the number of items. To avoid inverting an N+3n matrix, the procedure is broken into stages with each stage consisting of two parts. In the first part, the item parameters are held fixed and new abilities are estimated. In the second part, the abilities are held fixed and new item parameters estimated. Originally, the program did a straight run to convergence, where the stages were repeated until some convergence criterion was met. This was a very slow procedure that sometimes failed to converge. Over a period of ten years, the procedure was refined until an estimation procedure, called the automatic 4-step procedure, was finalized in 1976.

In a research study in 1989, Martha Stocking discovered a problem with this procedure (Stocking, 1989). For four sets of artificial data, the program produced different item parameter estimates when the true item discriminations were used as starting values than when the default starting value of one was used. However, when the original run to convergence procedure was used, the two different sets of starting values converged to the same parameter estimates. This paper explores several methods of correcting

this problem without resurrecting the problems that the 4-step procedure was designed to prevent.

The methods tried are

1. Computing the initial a parameter estimates from the conventional item statistics, the r biserial and the proportion correct.
2. Running the 4-step estimation procedure twice, for eight steps.
3. Running the 4-step estimation procedure with a prior on the item discrimination parameter.
4. Running the estimation procedure to convergence with a prior on the item discrimination parameter.
5. Running the estimation procedure to convergence but speeding the convergence by extrapolating the item parameter estimates.

6 and 7. Grouping the abilities for a short initial run to a loose convergence criterion to get initial item parameter estimates, and then running the 4-step procedure. Two methods of grouping were tried.

These seven new methods plus the current automatic 4-step procedure and the original run to convergence method were evaluated using artificial data.

Another issue that has concerned users of LOGIST is the way the program estimates the lower asymptote parameter when the item response function becomes asymptotic in a region of the ability distribution where there are few or no examinees. Occasionally when a group of items is calibrated in two separate LOGIST runs, each run containing other items, some lower asymptotes will be fixed at a common c value in one run but will be estimated in the other run. Although in both runs the estimated item response function will

usually fit the data well, the item parameter estimates are sometimes quite different. In addition, the estimated common c may be different for the two runs because the group of items used to estimate the common c will be different. This study also explored, using artificial and real data, whether a prior on the c's improves the consistency of the item parameter estimates for the items calibrated in two different LOGIST runs. Putting a prior on c is an eighth new method of estimating the parameters.

The Three Parameter Logistic Model

The item response model used by LOGIST is the Birnbaum three-parameter logistic, (3-PL), model. For a dichotomous item, the item response function, IRF, is the probability of a correct response to an item and is a function of the examinee's ability, θ , and three item parameters describing the item. The formula for the probability, $P_i(\theta)$, is

$$P_i(\theta) = c_i + (1-c_i) / (1+e^{-1.7a_i(\theta-b_i)}) \qquad (1)$$

where

a_i is the item discrimination index which is proportional to the slope of $P_i(\theta)$ at the point of inflection.

b_i is the item difficulty which is the point of inflection on the θ metric.

c_i is the lower asymptote of $P_i(\theta)$.

Development of the Current 4-Step Estimation Procedure

This 3-PL model has one parameter per examinee and three parameters per item, for a total of N + 3n parameters where N is the number of examinees and n is the number of items. Using joint maximum likelihood, LOGIST estimates the item and examinee parameters that simultaneously maximize the joint likelihood function modified for omits

$$L=\prod_{k=1}^{N}\prod_{i=1}^{n_k} P_i(\theta_k)^{v_{ik}} Q_i(\theta_k)^{(1-v_{ik})} \qquad (2)$$

where v_{ik} is 0 if item i was answered incorrectly by examinee k, 1 if answered correctly, and 1/A if omitted. A is the number of response alternatives. n_k is the number of items that examinee k reached. Since the likelihood function modified for omits is no longer strictly a likelihood function in the usual sense, it will be referred to hereafter as the criterion function. (It can, however, be described as a "limited-information likelihood function" which inherits properties associated with likelihood functions; see Mislevy and Wu, 1988). The parameters are estimated by setting the first derivatives of the log of the criterion function with respect to the unknown parameters to zero and solving these equations using Newton's method. Newton's method requires some initial starting values for the parameters and iteratively corrects these values to obtain the solution to the joint likelihood equations. To solve for the corrections to all of these unknowns at once would require inverting an N + 3n matrix. Since this is beyond the capacity of most computers, solving for the corrections is split into two parts. One part solves for the corrections to the ability estimates. The other part solves for the corrections to the item parameter estimates. These two parts

together are called a stage. In a straight run to convergence, stages are repeated until the percent change in the criterion function between two successive stages is less than a user specified criterion. The solution has two indeterminant values, the origin and scale of the ability metric. LOGIST handles this by standardizing the abilities to a robust mean of zero and a standard deviation of one. The robust mean and standard deviation are computed using Tukey's biweight method. (Mosteller & Tukey, 1977). The standardization is done between the ability estimation part and the item estimation part.

A lot of fine tuning was required to overcome some of the problems encountered with this procedure. One major problem was that, for some data, a few of the a's would tend towards infinity. To prevent this, an upper bound was placed on the a's. When the abilities were rescaled, this upper limit on the a's also had to be rescaled in the same way as the a's were rescaled to avoid a decrease in the criterion function when the rescaled a's at the maximum tried to exceed the maximum and were reset to the maximum. To understand how this can create problems, suppose the procedure is nearly converged with some a's at the upper limit. When the abilities are estimated, the items with a at the maximum will be fitted better if the abilities are a little more spread out. The standardization then pulls the abilities back in, but also raises the maximum a. Since the a's at the maximum want to be higher anyway, they are becoming higher by this effect on the abilities. The criterion function is increasing, but very slowly. If there are several a's at the maximum, the percent change in the criterion between successive stages can take a long time to become less than the convergence criterion.

Meanwhile, the a's at the maximum are increasing to a higher value than is wanted or reasonable.

To control this problem, Frederic Lord devised what has been called the automatic, or 4-step, procedure used in LOGIST since 1976. In this procedure, the a's and the abilities only interact twice. In step 1, the a's and c's are held fixed, while the abilities and b's are estimated until the percent change in the criterion function between two successive stages is less than a loose convergence criterion. In step 2, the abilities are held fixed, and the a's, b's, and c's are estimated. In the third step, again the a's and c's are held fixed, and the abilities and b's are estimated. In the fourth step, the abilities are again held fixed, and the item parameters are estimated. The convergence criterion is reduced for each step, starting with 200% for step 1 to .2% for step 4. In steps 1 and 3, the robust standardization is used. In steps 2 and 4, the standardization sets a truncated mean and standard deviation of the abilities to zero and one respectively. This method was tried on some artificial data and produced acceptable results. The important point of this procedure is that the abilities and the a's only interact twice. The a's aren't given a chance to increase without limit and the maximum a is not changed by the standardization. The initial a's are set to a constant. Different starting values for the constant converged to the same final parameter estimates.

This method seemed satisfactory until the Stocking study found problems using some extreme datasets. For some of these datasets, the 4-step procedure converged to different parameter estimates depending upon different starting values for the item parameters. In particular, she discovered that setting

the initial a's to the true values gave different and better results than using the default value of one for all of the a's. Since LOGIST isn't run to convergence, it is not surprising that starting at the true values gives better results. However, when LOGIST was run to convergence, the parameter estimates obtained with the initial a's set to the true a's agreed with the parameter estimates obtained with the initial a's set to a constant. The difference between the estimates for the automatic procedure with the default starting value for the a's and the estimates for the run to convergence is sufficient to warrant investigating ways to improve the current LOGIST procedure.

Another problem with the estimation procedure has been the difficulty of estimating the lower asymptote, c, for easy items or not very discriminating items where c is poorly determined because there are few to no examinees in the region where the IRF becomes asymptotic. In the development of LOGIST, several methods were tried to obtain reasonable estimates of c for these items. The method implemented was to fix the c's at a common c value for items where b-(2/a) is less than some cut-off criterion and estimate a common c for all items with c fixed. The value of b-(2/a) is the ability where the item response function approaches the lower asymptote. The common c value is estimated using Newton's method but only in the second step of the 4-step estimation procedure. Thus the common c value depends on the other items in the run that had their c's fixed at the common value. There are two problems with this procedure that are noticeable when two separate calibrations have items in common and a comparison is made of the estimated parameters for the common items. The problems are: 1) the common c value depends on the other

items fixed at the common value and may be quite different for two different runs; 2) the criterion value for determining whether c is fixed or not creates a discontinuity. In one run an item may have its c fixed, and in another, because the b-(2/a) happens to be slightly higher, the c may be estimated for this item. The fit of the item response curve to the data may be approximately the same in both cases although the parameter estimates may differ.

Methods

Seven methods were tried to improve the LOGIST 4-step procedure results without the expense of running to convergence. These methods are outlined in the introduction on page 5. The first three methods tried to improve the automatic procedure. These methods were: 1) computing the initial a parameter estimates from the conventional item statistics; 2) running the four step procedure for eight steps; and 3) putting a log normal prior on the a's and running the 4-step procedure. When none of these gave as good results as running to convergence, several modifications were tried to increase the speed of running to convergence. These additional methods were: 4) adding a log normal prior to the a's and running to convergence; 5) adding extrapolation to speed up the run to convergence; and 6) & 7) obtaining initial item parameter estimates by running to convergence with the examinees grouped into a small number of groups and then running the 4-step procedure. Finally, an eighth method put a prior on c to improve the poorly estimated c's. Each procedure will be discussed in detail.

1) Using Conventional Item Statistics to Get Initial a Estimates.

Stocking showed in her study that using the true item discrimination values as starting values for the a's gave better parameter estimates than using the default value of one for the a's. This suggests that better initial starting values for the a's would improve the final item parameter estimates. Lord (1980) gives approximations to the item discrimination and difficulty parameters, provided c is 0, that are computed from the r-biserial and the observed proportion correct. Schmidt (1977) gives Urry's modifications to these equations to correct for guessing. These formulas hold only if the unit of measurement for θ has been chosen so that the mean of θ is 0 and the standard deviation is 1 and θ is normally distributed in the group tested. In addition, the approximations can fall short of accuracy when the test score x and θ have differently shaped distributions.

The initial estimate for a is given by the following formula

$$a_i = \frac{\rho_{i\theta}}{\sqrt{1-\rho_{i\theta}^2}} \tag{3}$$

where

$$\rho_{i\theta} = \frac{\rho'_{i\theta}\sqrt{P'_i Q'_i}}{(1-c_i)\phi(\gamma_i)} \tag{4}$$

$\rho'_{i\theta}$ is the point biserial correlation between the binary item score and the latent trait, θ . The point biserial is attenuated by guessing. To the extent that the number-right score x is a measure of ability θ, ρ_{ix}, the

product-moment or point-biserial correlation between item score and x, can be used as an approximation to $\rho'_{i\theta}$

$\rho_{i\theta}$ is the correlation between the normally distributed dimension underlying the item and the latent trait θ .

$\phi(\gamma_i)$ is the ordinate at γ_i that cuts off the area P_i of the standardized normal curve.

P'_i is the observed percent correct.

$P_i=\frac{P'_i-c_i}{1-c_i}$ is the observed percent correct adjusted for guessing.

An upper bound of 1 was placed on the computation of the biserial correlation. In addition, the initial a estimates were not allowed to exceed the maximum value for a specified by the user.

2) Eight-Step Procedure

Since running the program to convergence gives the same results regardless of the starting value for the a's, will simply running the 4-step procedure for eight steps give acceptable results? Instead of repeating the 4-step procedure twice, steps 1 and 2 were repeated three times and then steps 3 and 4 were executed. This was done to avoid running to the tight convergence criterion of steps 3 and 4 in the middle of the estimation procedure.

3 & 4) Log normal prior on a for the 4-step and the Run to Convergence.

In Stocking's study, the problem seemed to be primarily with the estimation of the a's. Consequently, a prior on the a's might improve the estimation of the a's in the 4-step procedure. The prior tried was a

'floating' log normal prior as is used in the BILOG estimation program (Mislevy and Bock, 1983). The parameters of the prior are μ_a and σ_a , the location and dispersion. The σ_a can be specified by the user or set to a default value of .5. μ_a is the mean of the log a's and is recomputed at the end of each stage or it can be specified by the user and fixed. The criterion function in equation 2 becomes

$$L^* = \left(\frac{1}{\sigma_a\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln a - \mu_a}{\sigma_a}\right)^2}\right) L \qquad (5)$$

The μ_a is adjusted by the standardization of the abilities at the end of each stage. This method was tried with the 4-step procedure and with the run to convergence method.

5) Convergence method with extrapolation

As this study progressed it became obvious that running to convergence produced the best results in terms of reproducing the true parameters. However, running to convergence has always been a slow and expensive procedure. In addition, there is sometimes the problem of a's becoming unreasonably large in running to complete convergence. Consequently, several modifications of running to convergence were tried. The first was using extrapolation on the item parameters to speed convergence. The same extrapolation method was used as had previously been used in the 4-step procedure on groups of b's, only now it was applied individually to the a's and c's as well as to the b's. A logit transformation was done on the c's and a log normal transformation was done on the a's before extrapolation. After

extrapolation, the parameters were transformed back. No extrapolation was attempted unless the absolute change in the item parameter was monotonically decreasing. Each item parameter was extrapolated independently of the other parameters for that item. An absolute limit was put on the amount any parameter could change by extrapolation. In addition, no parameter could change by more than nine times the amount that it changed in the previous stage. Extrapolation was done every four stages after the maximum difference between the item response functions between two successive stages had become less than .01 and, if no prior was placed on c, the common c value had converged.

6 and 7) Convergence Method with Grouping

Two other methods were tried that involved grouping the abilities after the first estimation of abilities and running to convergence using the grouped abilities. The estimated item parameters were then used as initial starting values for the 4-step procedure. The reason for running the 4-step procedure after the grouped run to convergence is that the grouping tends to produce slightly underestimated a's. Two different groupings were tried. One grouped the examinees into 100 centiles with an equal number of examinees in each group. This results in groups with a narrow ability range in the middle of the ability distribution but with a wide ability range in the tails of the distribution. The other grouped the examinees into 100 equally spaced groups between -3 and 3 on the ability scale. Ability estimates outside of this range were put in the appropriate extreme group. This grouping allows for more groups in the tails of the ability distribution.

8) Prior on c

For some items, the item response function becomes asymptotic in a region of the ability distribution where there are few or no examinees. The ability level where the IRF becomes asymptotic is approximately equal to b-(2/a). After the c's have been estimated for two stages, for items with b-(2/a) less than some specified ability level below which there are few examinees, LOGIST fixes the c at a common c value that is then estimated by maximum likelihood. Thus, for items with c fixed, the c depends on the other items that are also fixed.

Another method of controlling the estimation of c for items where there is little information about c contained in the data is to adopt a Bayesian approach of controlling c by the imposition of a prior distribution. This prior distribution formally incorporates our beliefs about reasonable values of the c parameter. The prior distribution used was a beta function as implemented in BILOG (Mislevy and Bock, 1983). The beta function is

$$\frac{1}{B(\alpha,\beta)} c^{\alpha-1}(1-c)^{\beta-1} \qquad (6)$$

The parameters α and β are determined as follows. The mean of the beta distribution is set equal to the mean of the c's, $\bar{c}$. Thus,

$$\frac{\alpha}{\alpha+\beta}=\bar{c} \qquad (7)$$

The weight that is given to the prior can be expressed in terms of the number in a hypothetical group of low ability examinees, N_{-} .

$$\alpha+\beta=(N_{-\infty}+2) \tag{8}$$

Solving these two equations for alpha and beta gives

$$\alpha=(N_{-\infty}+2)\,\bar{c} \tag{9}$$

$$\beta=(N_{-\infty}+2)\,(1-\bar{c}) \tag{10}$$

$N_{-\infty}$ controls how tight or loose the prior is. For future reference, define $N^{*}_{-\infty}=N_{-\infty}+2$. $\bar{c}$ can be computed at the end of each stage or fixed at some user specified value. If it is computed, then the poorly determined c's will depend on the c's of the other items included in the calibration. If $\bar{c}$ is fixed, the poorly determined c's will depend on the value assigned to it. For any item, the estimated c will depend upon the prior to the degree that there is little information in the data with which to estimate c.

Convergence Criterion

It was necessary to change the convergence criterion twice during these runs for the methods that involved running to convergence. The original convergence criterion of some minimum percent change in the criterion function created problems. For three of the datasets used in Stocking's study, in the straight run to convergence after forty stages the criterion function was still changing by more than three percent, and the decrease in the percent change between stages was very small. For example, for one of the datasets the percent change in the criterion function for stages 48 to 54 were 3.61, 3.51, 3.44, 3.28, 3.20, 3.08, 3.04. The reason for this extremely slow convergence is that there were several items with a at the maximum a. The

standardization would increase the maximum a by a small amount. The a's at the maximum would then increase to the revised amount, and so the criterion function would increase by a small amount. This problem was not removed by putting a prior on a. For the same test with a prior on a, the percent change in the criterion function was 1.46, 1.34, 1.22, 1.14, 1.13 for stages 48 to 54.

Although the criterion function was increasing by more than the convergence criterion, the estimated item response functions were changing very little in the region of ability where the examinees were located. Consequently, the convergence was redefined to be the maximum change between the fitted IRF's within a stage. This change between the fitted IRF's was computed for theta values of -2, -1, 0, 1, and 2. The procedure was considered converged when the maximum difference between the estimated IRF's was less than .0005. A couple of datasets had problems with this criterion as the change within a stage would compensate for the adjustment by standardization so that the change across stages would be extremely small but the change within a stage would be larger than the convergence criterion. The final convergence definition chosen was that the maximum difference in the IRF's across two stages was less than .0005.

Data

Since the problems with LOGIST were discovered in Stocking's study, all methods used in this paper were tried on the four sets of artificial data used in that study. Each test contained 100 5-choice items and were taken by 3000 examinees. The true c's for all items were set to .15. This value was chosen

based on the observation that in practice c's are usually estimated as smaller than the probability of a correct answer based on random guessing. The true b's were chosen randomly from a rectangular distribution with a range of -2.5 to 2.5. The true a's were chosen to be within the range of .5 to 2.5, but with different correlations with the true b's for each test.

These correlations varied as follows:

1) For test S1, the population correlation between true a and true b was +.8, the sample correlation was +.76.

2) For test S2, the population correlation was -.8, the sample correlation was -.76.

3) For test S3, items with difficulty less than zero had a population correlation with item discrimination of -.8. Items with difficulty greater than zero had a population correlation with item discrimination of +.8. The overall sample correlation was .08.

4) For test S4, items with difficulty less than zero had a population correlation with discrimination of +.8, and items with difficulty greater than zero had a population correlation with discrimination of -.8. The overall sample correlation was .00.

These correspond to the datasets labeled C1 through C4, respectively, in the Stocking study.

A calibration sample of 3000 simulees were chosen from a rectangular distribution of true ability from -2.5 to 2.5. The population mean and standard deviation were 0 and 1.44, the sample mean and standard deviation were -.01 and 1.46. A separate criterion or cross-validation sample of 3000 simulees was chosen from the same distribution and had a sample mean and

standard deviation of -.03 and 1.45. Item response data for each test were generated for each simulee in each sample. For the calibration sample, these data were used to obtain the item and ability parameter estimates from LOGIST. For the cross-validation sample, these data were used to obtain only ability estimates using the item parameters estimated in the LOGIST run on the calibration sample.

In addition to this data, four more sets of data were generated to simulate real tests by first calibrating actual test data on LOGIST6 (Wingersky, et al, 1988) and then using the estimated parameters as the true parameters to generate item responses. Since there were actual responses associated with the true abilities, items not reached by the real examinee were considered not reached for the simulated examinee. Since LOGIST ignores 'not reached' items, degradation in the calibrations due to not reached items resulting in a shorter test will appear in the results. The total sample of examinees was split into two random halves. One half was used to generate the item responses for the calibration of the items, the other half was used to generate the item responses for the validation sample.

The four real tests used were:

1) R1, a GRE quantitative test containing 60 items and taken by 2264 examinees.

2) R2, an SAT math test plus equating section containing a total of 85 items and taken by 3015 examinees.

3) R3, an SAT verbal test plus equating section containing a total of 130 items and taken by 2744 examinees.

4) R4, a TOEFL section containing 58 items and taken by 2915 examinees.

In comparison to tests S1 to S4, the correlations between the true a and b parameters were .63 for R1, .62 for R2, .01 for R3, and .37 for R4.

Analysis

Each of the eight datasets were calibrated using the eight new procedures, the 4-step procedure, and the run to convergence. The estimated parameters were transformed to the scale of the true parameters using the method which minimizes the squared difference between the two test characteristic curves, as described in Stocking and Lord (1983). The transformed item parameter estimates were then used to estimate abilities for the validation samples.

Because of the numerous methods involved, the following abbreviations were used for the different methods. The 4-Step procedures are referred to as the automatic procedures, as the default automatically used by the program was the 4-Step procedure.

Abbreviation or code	Method
4	Automatic 4-Step procedure currently used in LOGIST.
r	Automatic 4-Step procedure where the initial a's are computed from the r-biserials and observed proportion correct.
A	Automatic 4-Step procedure with a log normal prior on the a's.

8 Automatic 8-Step procedure. This is a procedure with steps 1 and 2 of the 4-Step procedure repeated three times and then steps 3 and 4 of the automatic procedure are done.

b Run to convergence with a log normal prior on a.

C Run to convergence.

E Run to convergence with extrapolation of the item parameters.

= First a run to convergence with extrapolation is done with the abilities grouped into 100 intervals of equal width between -3 and 3 on the ability scale. This run is then followed by the 4-step procedure.

% First a run to convergence with extrapolation is done with the abilities grouped into 100 intervals where the abilities are sorted and grouped into centiles with one percent of the examinees in each group. This run is then followed by the 4-step procedure.

P Prior on c was added to the run to convergence with the abilities grouped into 100 intervals of equal width between -3 and 3, followed by the 4-step procedure.

The estimated item parameters were evaluated by 1) comparing the estimated parameters to the true parameters, 2) comparing the fitted item response functions to the true item response functions, and 3) comparing abilities estimated with the estimated item parameters to abilities estimated with the true item parameters on a separate validation sample.

The following statistics were computed for comparing each estimated item parameter to the corresponding true item parameter for the a's, b's and c's. 1) The correlation between the estimated and the true parameters. For the c's, no correlation was computed for the S tests since all of the true c's were equal. 2) The bias, which is the estimated parameter minus the true parameter, averaged over all items. 3) The root mean square error, RMSE, which is the square root of the average squared differences between the estimated and the true parameters. These statistics for all of the methods for all of the tests are given in Table 1.

Insert Table 1 about here

Since different combinations of a, b and c can produce similar item response functions in the range where there are examinee abilities, comparing the estimated and true item response functions in this range may be more meaningful than comparing the individual estimated and true tem parameters. The following statistics comparing the estimated item response functions to the true item response functions were computed: 1) ASD, the unweighted average signed deviation between the two curves, 2) RMSD, the square root of the average unweighted squared deviation, 3) BIAS, the average weighted signed

deviation and 4) WRMSD, the square root of the average weighted squared deviation. ASD and RMSD were computed over equally spaced abilities between -2.5 and 2.5 for the S tests and between -3 and 3 for the R tests. The smaller range was used for the S tests since the true abilities spanned only this range. BIAS and WRMSD were computed over equally spaced abilities between -3 and 3 and were weighted by the expected number of examinees for the appropriate segment of the ability distribution assuming the abilities to be normally distributed with a mean of 0 and a standard deviation of 1. This gives four statistics for each item. The individual statistics were plotted by item, where the items were ordered by true difficulty, to see whether systematic differences occurred at different ability levels. These plots were too numerous to include in this paper, but are available from the author.

The mean and standard deviation of each of these statistics were computed for each estimation method for each test. Figure 1 contains plots of the means for the automatic methods plus the run to convergence method. Figure 10 contains the plots of the means for the convergence methods. The scale for the y-axis in Figure 10 is half that in Figure 1. These figures contain four plots, one plot for each statistic. Each plot contains the means for each test with a dotted line separating the tests and the test identification written below the x-axis. A point is plotted as a box containing the code for the method with lines extending one standard error of the mean from the center of the box.

Another criterion for evaluating the methods is to compare abilities estimated with the estimated item parameters to abilities estimated with the true item parameters. Comparing these two estimated abilities removes from

the comparison any error caused by the maximum likelihood estimation. The abilities were estimated on separate validation samples. The residuals between these two estimated abilities were grouped on the true abilities into intervals of .2. The median residual for the group and the lower and upper limits of a non-parametric two tailed 5% confidence band around the median, based on a method by David (1981), were plotted.

Figures 2 through 9 contain the plots of the medians for the automatic methods and for the run to convergence with each test in a separate figure. Figures 11 through 18 contain the plots for the convergence methods. The scale for the y-axis in Figures 11 through 18 is half the scale for the y-axis in Figures 2 through 9. For each ability group, the medians for the methods being compared are plotted in a separate position on the x-axis. The symbol for the method plotted is printed just below the x-axis. The ability groups are separated by a dotted line with the true ability printed beneath each group. The number of simulees in each group (N) is also printed. Due to this spreading of the points, the plot is broken into three sections. The top section contains the low abilities, the middle section contains the middle abilities, and the bottom section the high abilities. For the S tests, the groups were of uniform size, but for the R test the groups in the extremes were quite small. No confidence limits were computed if there were less than four examinees in a group.

Insert Figures 1 to 18 about here

Results

The discussion of the results will be broken into two sections. First the automatic methods will be compared to each other and to the run to convergence. Then the convergence methods will be compared to the run to convergence. The run to convergence was used as the criterion to judge the other methods since, for the most part, this method gave the best parameter estimates. Each method will be evaluated separately.

Method 4, the current four step procedure.

How bad is the current 4-step procedure (Method 4) that has been used extensively since 1976 compared to running to convergence (Method C)?

The statistics for comparing the item parameters for these two methods are given in Table 1. For the S tests, Method C gave better results than Method 4 with a few exceptions. For the R tests, the statistics for the two methods were nearly the same. Method C gave higher correlations for the a's for two of the R tests, but slightly lower correlations for the other two. The largest differences between the two methods were for R1. In looking at these statistics, Method C is not much better than Method 4 for tests that simulate real data, the R tests.

Figure 1 shows that Method C gives better fitting item response functions than Method 4. Except for the ASD, the difference between Method C and Method 4 was a lot less for the R tests than for the S tests. That the difference between the two methods was greater for the ASD than for the BIAS for the R tests indicates that Method 4 had problems in the extremes of the ability distribution. The plots of the statistics ordered by difficulty confirm this.

This is also confirmed by Figures 2 through 9, comparing the estimated abilities. Method 4 is the first procedure in a group, Method C the last. For the S tests, Method C gave definitely better results. For the R tests, the results for Method C were about the same or only slightly better than the results for Method 4. Rarely, as in R2, for abilities at -1.5 and -1.3, Method C did slightly worse than Method 4. For the R tests, throughout the middle ability range between -1 and 1, Method C and Method 4 agreed very well. Overall, the Method 4 results were closer to the Method C results for the R tests than for the S tests.

Method r, using conventional item statistics.

Does computing the initial a estimates from the conventional item statistics (Method r) give better estimated parameters than using the default values of 1 for the initial estimates?

In the comparison of the estimated item parameters given in Table 1, Method r gave lower correlations than Method 4 for the S tests. However, for the R tests the correlations were slightly higher. For the bias and RMSE, Method r gave higher values than Method 4 for the S tests, but lower values for the R tests. The detailed plots by item in the supplement to this paper show that Method r gave poorer fitting IRF's for the easy items than the other automatic methods. In Figure 1, comparing the item response functions, except for tests S3 and S4, Method r produced better fitting IRF's than the other automatic methods. For the R tests, Method r gave almost as good results as Method C.

In the comparison of ability estimates, for the S tests, Method r gave very poor lower ability estimates compared to the other automatic methods. However, for the R tests, again Method r gave better ability estimates than the other automatic methods, but not as good as Method C. The reason for this discrepancy between the results for the S tests and the results for the R tests is that using the conventional item statistics to approximate the item characteristic curve parameters assumes that the abilities are normally distributed. This was approximately true for the R tests, but was not true for the S tests, where the abilities have a rectangular distribution.

Although Method r gave better parameter estimates than Method 4 for the R tests, it gave worse parameter estimates for the S tests.

Method A, prior on a

Stocking (1989) found that the main problem in the estimation procedure seemed to be with the a parameters. Would the estimated parameters behave better if the a's were controlled by a log normal prior with the mean estimated and the variance fixed? The mean of the log normal prior was set to the mean of the log of the a's at the end of each stage. The standard deviation was fixed at .5 at first. On the first dataset calibrated, several a's tried to go to infinity, so the value was reduced to .4. This value kept all of the a's within bounds for all of the tests except R4. For this test, a value of .3 was necessary to prevent any a's from approaching infinity.

In comparing the item parameter estimates, the item response functions and the estimated abilities, the Method A results were nearly identical to the

results for Method 4. The lack of improvement in results does not warrant the additional assumptions and subjective decisions required by the prior.

Method 8, eight steps.

Running the four step procedure for four more steps (Method 8) should give better item parameter estimates without the cost of running to convergence. Are the improved item parameter estimates close to the Method C estimates? Table 1 shows that the item parameter estimates were improved over Method 4 estimates. Figure 1 shows that Method 8 gave slightly improved results over Method 4, but still not nearly as good as Method C for the S tests.

In conclusion, Method C generally produced better item parameter estimates than any of the automatic methods. The improvement in the parameter estimates was greater for the S tests than for the R tests.

Convergence Methods

While running LOGIST to convergence gave better results than any of the automatic procedures, it was approximately twice as slow as the 4-step procedure and had the problem of some of the a's tending towards infinity. A comparison of the execution times for Method 4 and the convergence methods is given in Table 2. Table 2a gives the execution times. Table 2b expresses the times in terms of the percent of the time for Method 4. Attempts to increase the speed of Method C included extrapolation, grouping the abilities, and putting a prior on the a's. In total, there are six convergence methods, Methods C, E (extrapolation), = (grouping with equal interval width), %

(grouping into percentiles), b (prior on a), and P (prior on c). Methods C, b, E, = and % will be evaluated in this section. Method P will be evaluated in the following section.

Insert Table 2 about here

The runs to convergence for the S tests both with and without a prior on the a's used the percent change in the criterion function as the convergence criterion. The runs to convergence with and without a prior on the a's for the R tests were run with the convergence defined by the maximum change within a stage between the fitted IRF's computed for abilities at -2, -1, 0, 1, and 2. The run to convergence with extrapolation and the two grouping methods were run with convergence defined as the maximum difference between the estimated IRF's from one stage to the next. The convergence criterion was .0005.

Method E, extrapolation of item parameters.

Extrapolating the item parameters (Method E) improved the speed of convergence over Method C and produced nearly identical results to those for Method C. However, for most of the tests, the procedure was slower than Method 4. This is shown in Table 2.

Method b, running to convergence with a prior on a.

Running to convergence with a log normal prior on the a's (Method b) did not produce as good item parameter estimates as those from Method C. This is

shown in Table 1 and Figure 10. The ability estimates were also not quite as good. The prior did not improve the speed of convergence.

Methods = and %, grouping examinees.

These two grouping methods get initial values for the item parameters by grouping the abilities into coarse groups and running to convergence . The four step procedure is then run to correct for underestimation of the a's produced by the grouping. First these methods will be compared to Method C and then they will be compared to each other.

In comparing the item parameter estimates, grouping gave generally slightly higher correlations between estimated and true parameters. Grouping slightly increased the bias in the a's for the S tests and slightly decreased the bias in the a's for the R tests. Grouping decreased the RMSE for the b's and c's. In comparing the item response functions in Figure 10, for most of the tests grouping improved the results compared to Method C. S2 was an exception; grouping gave slightly worse results. Compared to Method C, grouping improved the lower ability estimates for tests S2 and S3, markedly. However, for S2, grouping did not produce high ability estimates that were as good as those produced by Method C. In comparing Method = to Method %, Method = gave slightly better parameter estimates in most cases.

For five of the tests, grouping ran in considerable less time than Method 4, as is shown in Table 2. However, for one of the tests, grouping increased the time by 26 percent.

Since Method = produced nearly as good or better results than Method C in less computer time, Method = has been incorporated into LOGIST. One can

think of this method as running a step 0, which is a run to convergence with the examinees grouped into 100 intervals from -3 to 3 on the basis of the first maximum likelihood ability estimate. This step 0 gives initial item parameter estimates for the 4-step procedure. However, this step 0 is not without its problems. There is still the problem with a's restricted to the maximum value slowly increasing because of the standardization and causing the step to take a long time to converge. Since step 0 is simply to get good initial parameter estimates, it has been restricted to forty stages in length. This was more than enough for all of the tests run in this study to converge.

Prior on c

Having chosen a method that gives good parameter estimates, (i.e. Method =), can the estimates for items where the IRF reaches a lower asymptote in a region of the ability distribution where there are few to no examinees be improved by putting a beta prior on the c's? For these items, at the end of the third stage of step 0, Method = fixes the c for the remainder of Step 0 at a common c value if b-(2/a) is less than some criterion. The c's for all items are estimated again in stages 2 and 3 of step 2. At the end of stage 3 of step 2, the c's for items with b-(2/a) less than the criterion are fixed at a common c. The criterion used in all of these runs was -4. The common c is estimated and depends on the other items with c fixed.

The program using Method = was modified to put a beta prior on c with the mean of the prior either estimated or held fixed. All datasets were run with a prior on the c with the mean of the prior estimated and with N_{-}^{*} at 20. N_{-}^{*} controls the amount that the prior controls the c estimates. A large value for N_{-}^{*} will tightly control the c's; a smaller value will give

the c's more freedom. In looking at four sets of real data, a value of 20 gave a standard deviation for the beta distribution close to the observed standard deviations of the estimated c's.

In Figure 10 containing the plots of the average IRF statistics, it can be seen that the prior on c (Method P) improved the estimated parameters for tests S2 and S3. For the other tests, particularly for R1, the estimated IRF's were worse.

In Figures 11 through 18, the residual ability plots, the point corresponding to a prior on c is the last point in each group. It should be compared with the point preceding it for Method = . For S1, the prior produced higher residuals for the lower abilities, although it reduced the residuals for the higher abilities. However for S2 and S3, the prior dramatically improved the residuals for the lower abilities and slightly improved the residuals for the higher abilities. For S4 the prior gave larger residuals for the lower abilities. For R1, the prior gave larger residuals throughout the ability range. For R2 and R3 the prior improved the ability estimates. However, for R4 the prior improved the estimates for the high abilities but not for the lower abilities.

Figures 19 to 26 contain the plots of the residuals for the item parameters and IRF statistics for Method P and Method =. All values are plotted against the true b-(2/a). The prior should improve the estimates for the lower values of b-(2/a). The results for no prior on c are plotted in the first column, for a prior on c in the second. The IRF statistics were included to show any effect that the different parameter estimates might have on them. For the S tests, the prior on c improves the estimates for the

poorly estimated c's but does not eliminate the scatter for the moderately poorly estimated c's. For R1, the prior on c improved the estimates for the items with poorly estimated c's but gave worse parameter estimates for some of the other items. For R2, it improved the c estimates for the really poorly estimated c's, but increased the scatter slightly for some of the moderately poorly estimated c's. For R4, the prior didn't do as well for the poorly estimated c's.

Insert Figures 19 through 26 about here

For the comparison of the estimated parameters to the true parameters in Table 1, compare the line Prior on c to the line immediately above it, Grouping, Equal Int. On the plus side, the prior on c: 1) improved the correlations for a,b, and c for all tests except R1 and R4, 2) reduced the bias and RMSE for a for all tests, 3) reduced the RMSE for b for all tests except R1 and R4 and 4) decreased the RMSE for c for all tests except S1, R1, and R4. On the negative side, the prior on c increased the magnitude of the bias in b for all tests except S4, R2, and R4 and increased the magnitude of the bias in c for all tests except S2, S3 and R2. Except for the bias in a, all of the statistics for R1 were worse with the prior on c than without.

To understand why the prior on c produced such poor results for R1, two more calibrations were done. In one calibration, $N^*_{..}$ was set to 40. This did not improve the results. In the other calibration, four items with the highest estimated c's were removed to see whether these were causing an unusual distribution of the c's. The items removed had c's greater than .27.

In the calibration with four item removed and $N^*_{..}$ equal to 20, the mean c went to the minimum of .05 and the procedure stopped. $N^*_{..}$ was increased to 40 and the calibration finished. Calibrating without the four items also failed to improve the results. However, these two calibrations enable one to look at the effect of two different $N^*_{..}$'s on the parameter estimates and the effect of removing a few items on the calibration of the remaining items.

The comparison of these parameters estimates are shown in Figure 27. The left column of plots compares the parameters estimated with $N^*_{..}$ at 20 to the parameters estimated with $N^*_{..}$ at 40. All points are plotted against the true b-(2/a). With $N^*_{..}$ at 40, the b's and the c's were higher for most of the items. The right column of plots compares the parameters estimated with the four items removed to the parameters estimated with all of the items. Both sets of parameters were calibrated with $N^*_{..}$ at 40. For most of the items, both the b's and the c's were higher when all of the items were calibrated. In contrast, Figure 28 contains the comparison of the parameters estimated with no prior on c with the four items removed to the parameters estimated with no prior on c with all items. The two sets of estimated parameters agree better than when the prior on c was used. It is important to note that if the mean of the prior on c were held fixed, removing the four items would not have as much effect on the other items.

Insert Figures 27 and 28 about here

Common Item comparison

Frequently the parameters for a block of items will be estimated in two separate calibrations containing other sets of items. It is not unusual for an item to have c estimated in one calibration and have the c fixed at the common c value in the other calibration if b-(2/a) is close to the criterion for fixing c. In addition, the common c value depends on all items with c fixed at this value and may be different for the two calibrations. A prior on the c will remove the discontinuity at the criterion for fixing c. However, if the mean of the prior on c is estimated, it depends on the items in the calibrations and may differ for the two calibrations. One way to avoid this problem is to fix the mean of the prior in one calibration to the estimated mean in the first calibration. This approach can also be applied to estimation without a prior on c by fixing the common c value in one calibration to the estimated common c value in the first calibration.

To evaluate whether the prior on c gave more consistent parameter estimates, four sets of data were calibrated, both with and without a prior on c. Each set were based on responses from two groups of examinees. One group took Form 1 of the test and an equating section. The other group took Form 2 and the same equating section. Two sets of real SAT data were used; a Verbal test and a Math test. In addition two sets of simulated data were generated. The true parameters for this simulated data were obtained from a concurrent calibration of the real data, where the two forms were calibrated in one LOGIST run. This put the true parameters for both forms on the same scale. The real data will be referred to as Math and Verbal. The artificial data generated from the calibration of the real math data will be referred to as

R2. The artificial data generated from the calibration of the real verbal data will be referred to as R3. Form 1 for these two tests is the same as the R2 and R3 tests discussed in the previous sections.

Each form was calibrated separately. For the artificial data, the estimated parameters were transformed to the scale of the true parameters using all items and the transformation procedure developed by Stocking and Lord (1983). For the real data, Form 2 was transformed to the scale of Form 1 using the common items with the Stocking and Lord transformation method.

Form 1 was calibrated with a prior on c and without a prior on c. Form 2 was calibrated in four ways: 1) without a prior on c with the common c estimated, 2) without a prior on c with the common c fixed at the common c value estimated for Form 1, 3) with a prior on c with the mean of the prior estimated, and 4) with a prior on c with the mean of the prior fixed at the mean estimated for Form 1.

Table 3 contains the statistics comparing the estimated item parameters to the true parameters for the common items for the artificial data for Form 1 and all calibrations of Form 2. "Fix comc" indicates the calibration with no prior on c where the common c value was fixed at the value estimated in the Form 1 calibration. The prior improved the results for R2 but not for R3.

Insert Table 3 about here

Table 4 contains the statistics comparing the item parameter estimates between Form 1 and Form 2 for the common items. For R2, the prior improved the correlation, bias and RMSE for b and c. For R3, the prior degraded the

correlation and bias for c. For Math and Verbal, the results with and without the prior were nearly the same.

Insert Table 4 about here

Fixing the mean of the prior slightly improved the c estimates with the prior for R3 and Verbal but made almost no difference for the other tests. With no prior on c, fixing the common c made almost no difference in the estimates for the two calibrations. In replicating item parameters in two calibrations. the prior either had no effect or gave only a slight improvement.

The prior on the c improved the results for most of the tests. However, for R1 the prior on c produced parameter estimates that were not as good as the estimates produced without a prior. While adding the prior solves some problems, others are created. Without a prior on c, only the c's fixed at the common c were directly affected by the other items in the calibration and then only by those other items also fixed. With a prior on c and the mean c estimated, the estimated c's for most of the items are affected by the other items. The prior on c does not remove the scatter for moderately poorly estimated c's, where b-(2/a) is between -2 and -4, that occurs without a prior on c. Fixing the mean c would remove the effect of the particular group of items being calibrated on the estimates but requires some prior knowledge of an appropriate value for the mean c. The prior on c also requires specifying some value for $N_{..}^{*}$. Since the prior on c cannot be recommended without

reservations, a program option was added so that the user could choose whether or not to put a prior on c.

Comparison of Method = Estimates to Method 4 Estimates

The method that gave the best results in a reasonable amount of computer time is Method =. This method gives different item parameter estimates than the current method, Method 4. To give an idea of how different the estimates are, residuals from a comparison of the estimated parameters from the two methods are plotted in Figures 29 to 32 for the artificial data and in Figure 33 for the real data. The tests are plotted two tests per page. For each test, the residuals between the parameter estimates are plotted against the Method 4 estimated parameters. The residuals for c are plotted against the Method 4 b-(2/a). The bottom plot contains the test characteristic curves for the two methods and, for the artificial data, for the true parameters. The test characteristic curves are scaled to length one. The most noticeable difference is that the more discriminating items have higher a estimates when estimated using Method =. This makes the test characteristic curves steeper for the new method. The plot of the residuals for the b parameter indicates a nonlinear relationship between the parameters estimated with the new method and the parameters estimated with the 4-step procedure.

Insert Figures 29 through 33 about here

Conclusions

Stocking (1989) discovered that the automatic procedure used by LOGIST produced different item parameter estimates when the initial a's were set to the true values than when the default constant of one was used. However, when the automatic procedure was bypassed and LOGIST was run to convergence, the different initial a values converged to the same item parameter estimates. Running LOGIST to convergence is a time consuming and costly procedure. There is also a problem with a's tending towards infinity. This paper investigated possible revisions to the program to improve the final item parameter estimates without the cost of running to convergence.

The method that gave the best results in a reasonable amount of computer time was a method that obtained initial item parameter estimates by grouping the abilities into 100 groups between -3 and 3 and running to convergence. The automatic 4-step procedure was then run with these initial item parameter estimates. This method has been incorporated into LOGIST.

The new procedure gave item parameter estimates that are closer to the true values than the current 4-step method. However, there is a definite nonlinear relationship between the estimated item parameters for the two methods after the parameters have been linearly transformed to the same scale. Consequently, in an ongoing series of calibrations, switching to the new procedure from the old 4-step procedure will produce a discontinuity in the parameter estimates in the same manner as would be caused by switching from LOGIST to BILOG.

In addition, a prior on the c's was tried to see whether two problems with the current procedure could be removed. The problems occur when the same set of items is calibrated with different sets of other items in separate calibration runs. One problem is that the estimated common c depends on the items in the calibration with c fixed at the common c. The other problem is that an item may be fixed in one calibration because b-2/a was slightly lower than the cut-off criterion and estimated in another calibration because b-2/a was slightly higher than the cut-off criterion.

The addition of a prior on c improved the parameter estimates for most of the tests. However, for one of the tests, the prior made the estimated parameters worse. The beta prior on c with the mean of the prior estimated makes most of the c estimates dependent on the other items being calibrated. With no prior on c, only the c's fixed at a common c value are dependent on the other items and then only on the items with c also fixed at the common c. The prior on c only slightly improved the replicability of the estimated item parameters when items are calibrated with other tests in separate calibration runs. Since the results were not conclusive, an option was added so that a user could specify that the program estimate the c's with a prior on c with the mean of the prior either fixed or estimated or the program estimate c without a prior with either the common c value estimated or fixed at a value specified by the user. More experience needs to be acquired before a clear recommendation can be made about using a prior on c.

The new LOGIST is called LOGIST7.

References

David, H. A. (1981). Order statistics (2nd. ed.). New York: Wiley.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Mislevy, R. J., & Bock, R. D. (1983). BILOG: Item analysis and test scoring with binary logistic models (computer program). Mooresville, IN: Scientific Software, Inc.

Mislevy, R. J., & Wu, P-K (1988). Inferring examinee ability when some item responses are missing. (Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Mosteller, F., & Tukey, J. W. (1977). Data analysis and regression. Reading, MA: Addison-Wesley.

Schmidt, F. L. (1977). The Urry method of approximating the item parameters of latent trait theory. Educational and Psychological Measurement, 37,. 613-620.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Stocking, M. L. (1989) Empirical estimation errors in item response theory as a function of test properties (Research Report 89-5). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Patrick, R., and Lord, F. M. (1988). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Table 1

Summary Statistics Comparing Estimated Item Parameters to True Parameters

		Correlation			Bias			RMSE		
		a	b	c	a	b	c	a	b	c
S1										
4 - Step	4	0.802	0.996		-0.044	0.035	-0.024	0.237	0.147	0.050
4 Step, R-biserial	r	0.686	0.991		-0.090	0.043	0.032	0.321	0.277	0.066
Prior on a, Auto.	A	0.804	0.996		-0.052	0.038	-0.023	0.235	0.147	0.048
8 step	8	0.844	0.997		-0.013	0.021	-0.024	0.212	0.124	0.052
Prior on a, Conv.	b	0.872	0.998		0.070	0.017	0.005	0.237	0.093	0.042
Convergence	C	0.925	0.998		0.074	0.010	0.006	0.187	0.096	0.043
Conv. with Extrap.	e	0.925	0.998		0.074	0.010	0.006	0.187	0.096	0.043
Grouping, Centiles	%	0.935	0.998		0.074	0.002	0.001	0.177	0.082	0.031
Grouping, Equal Int.	=	0.937	0.998		0.074	0.002	0.001	0.176	0.080	0.029
Prior on c	P	0.935	0.999		0.063	-0.006	-0.010	0.171	0.077	0.031
S2										
4 - Step	4	0.753	0.995		0.116	-0.063	0.012	0.295	0.153	0.029
4 Step, R-biserial	r	0.691	0.985		-0.087	0.031	0.026	0.360	0.342	0.056
Prior on a, Auto.	A	0.749	0.995		0.113	-0.065	0.012	0.292	0.156	0.028
8 step	8	0.798	0.996		0.109	-0.052	0.013	0.272	0.140	0.036
Prior on a, Conv.	b	0.881	0.997		0.096	-0.027	0.011	0.217	0.104	0.061
Convergence	C	0.898	0.998		0.074	-0.020	0.009	0.197	0.101	0.061
Conv. with Extrap.	e	0.896	0.998		0.072	-0.020	0.009	0.197	0.102	0.061
Grouping, Centiles	%	0.890	0.998		0.096	-0.023	0.013	0.212	0.098	0.055
Grouping, Equal Int.	=	0.900	0.998		0.099	-0.023	0.007	0.207	0.096	0.057
Prior on c	P	0.930	0.999		0.089	-0.026	-0.004	0.178	0.077	0.030
S3										
4 - Step	4	0.759	0.995		-0.105	-0.014	0.021	0.289	0.197	0.066
4 Step, R-biserial	r	0.421	0.977		-0.180	-0.145	0.035	0.453	0.520	0.065
Prior on a, Auto.	A	0.759	0.995		-0.114	-0.014	0.021	0.292	0.199	0.064
8 step	8	0.836	0.996		-0.022	-0.019	0.018	0.232	0.155	0.059
Prior on a, Conv.	b	0.873	0.997		0.070	-0.001	0.029	0.237	0.128	0.093
Convergence	C	0.892	0.997		0.063	0.001	0.034	0.213	0.128	0.101
Conv. with Extrap.	e	0.893	0.997		0.055	0.001	0.034	0.206	0.128	0.101
Grouping, Centiles	%	0.911	0.998		0.083	0.002	0.022	0.214	0.101	0.079
Grouping, Equal Int.	=	0.914	0.998		0.091	-0.001	0.018	0.217	0.101	0.076
Prior on c	P	0.946	0.999		0.080	-0.017	-0.012	0.178	0.066	0.034
S4										
4 - Step	4	0.919	0.999		0.117	0.011	-0.010	0.199	0.081	0.032
4 Step, R-biserial	r	0.931	0.996		0.052	-0.020	0.025	0.205	0.140	0.050
Prior on a, Auto.	A	0.921	0.999		0.121	0.008	-0.011	0.197	0.082	0.035
8 step	8	0.938	0.999		0.082	0.017	-0.009	0.170	0.073	0.033
Prior on a, Conv.	b	0.948	0.999		0.090	0.015	0.009	0.169	0.082	0.052
Convergence	C	0.943	0.998		0.061	0.028	0.009	0.159	0.091	0.052
Conv. with Extrap.	e	0.942	0.998		0.062	0.027	0.009	0.160	0.090	0.052
Grouping, Centiles	%	0.954	0.999		0.072	0.020	0.006	0.153	0.066	0.029
Grouping, Equal Int.	=	0.954	0.999		0.073	0.019	0.005	0.154	0.067	0.029
Prior on c	P	0.954	0.999		0.059	0.008	-0.012	0.146	0.063	0.029

Table 1 (Cont'd)

Summary Statistics Comparing Estimated Item Parameters to True Parameters

		Correlation			Bias			RMSE		
		a	b	c	a	b	c	a	b	c
R1										
4 - Step	4	0.922	0.974	0.578	-0.011	-0.017	-0.027	0.124	0.282	0.116
4 Step, R-biserial	r	0.929	0.972	0.613	0.028	-0.009	0.006	0.140	0.300	0.103
Prior on a, Auto.	A	0.928	0.979	0.654	-0.016	0.011	-0.020	0.112	0.250	0.109
8 step	8	0.935	0.979	0.687	0.018	-0.001	-0.011	0.131	0.253	0.102
Prior on a, Conv.	b	0.923	0.975	0.658	0.043	0.034	0.013	0.151	0.280	0.109
Convergence	C	0.951	0.978	0.686	0.040	0.012	0.010	0.129	0.264	0.102
Conv. with Extrap.	e	0.952	0.978	0.686	0.040	0.012	0.010	0.128	0.264	0.102
Grouping, Centiles	%	0.944	0.978	0.708	0.044	0.009	0.011	0.142	0.264	0.097
Grouping, Equal Int.	=	0.944	0.978	0.698	0.043	0.012	0.012	0.141	0.260	0.098
Prior on c	P	0.920	0.951	0.530	-0.004	-0.119	-0.050	0.151	0.423	0.109
R2										
4 - Step	4	0.913	0.991	0.771	-0.038	0.014	-0.014	0.134	0.175	0.058
4 Step, R-biserial	r	0.924	0.992	0.799	-0.021	0.018	-0.003	0.122	0.158	0.053
Prior on a, Auto.	A	0.904	0.987	0.746	-0.043	0.030	-0.010	0.144	0.206	0.062
8 step	8	0.939	0.992	0.773	-0.009	0.001	-0.009	0.110	0.159	0.058
Prior on a, Conv.	b	0.939	0.989	0.733	0.010	0.032	0.014	0.110	0.187	0.065
Convergence	C	0.951	0.990	0.741	0.029	0.022	0.016	0.109	0.176	0.064
Conv. with Extrap.	e	0.951	0.990	0.741	0.031	0.021	0.017	0.110	0.177	0.064
Grouping, Centiles	%	0.949	0.990	0.743	0.028	0.024	0.018	0.110	0.174	0.063
Grouping, Equal Int.	=	0.949	0.990	0.761	0.025	0.020	0.014	0.109	0.174	0.060
Prior on c	P	0.951	0.992	0.779	0.004	-0.012	-0.007	0.100	0.161	0.055
R3										
4 - Step	4	0.888	0.988	0.777	-0.010	-0.006	-0.007	0.126	0.205	0.067
4 Step, R-biserial	r	0.886	0.988	0.771	0.003	-0.016	-0.006	0.128	0.208	0.068
Prior on a, Auto.	A	0.888	0.986	0.762	-0.012	-0.003	-0.005	0.124	0.208	0.069
8 step	8	0.884	0.987	0.772	0.001	-0.015	-0.007	0.130	0.212	0.069
Prior on a, Conv.	b	0.882	0.989	0.802	0.013	0.001	0.007	0.128	0.198	0.064
Convergence	C	0.876	0.987	0.760	0.018	-0.007	0.004	0.136	0.216	0.070
Conv. with Extrap.	e	0.875	0.987	0.760	0.020	-0.008	0.005	0.136	0.216	0.070
Grouping, Centiles	%	0.880	0.987	0.774	0.015	-0.013	0.003	0.134	0.214	0.068
Grouping, Equal Int.	=	0.882	0.987	0.773	0.009	-0.013	-0.001	0.132	0.212	0.068
Prior on c	P	0.888	0.990	0.786	-0.008	-0.015	-0.005	0.126	0.192	0.063
R4										
4 - Step	4	0.945	0.996	0.872	0.003	-0.014	-0.017	0.091	0.130	0.047
4 Step, R-biserial	r	0.934	0.995	0.886	0.051	-0.027	-0.004	0.138	0.129	0.039
Prior on a, Auto.	A	0.935	0.996	0.874	-0.006	-0.008	-0.018	0.090	0.125	0.046
8 step	8	0.952	0.995	0.860	0.025	-0.028	-0.017	0.108	0.142	0.051
Prior on a, Conv.	b	0.910	0.995	0.851	0.040	0.006	0.006	0.149	0.136	0.045
Convergence	C	0.939	0.994	0.856	0.064	-0.011	0.008	0.155	0.143	0.045
Conv. with Extrap.	e	0.939	0.994	0.856	0.061	-0.010	0.008	0.149	0.143	0.045
Grouping, Centiles	%	0.942	0.995	0.880	0.053	-0.019	0.005	0.137	0.135	0.041
Grouping, Equal Int.	=	0.942	0.995	0.880	0.050	-0.024	-0.004	0.136	0.134	0.042
Prior on c	P	0.946	0.995	0.829	0.029	-0.054	-0.024	0.120	0.153	0.051

Table 2a

Time in Seconds on an 80386 - 20 MHZ.

	4 Step	Convergence	Convergence Extrap.	Grouped Equal Int.	Grouped Centiles
S1	3477	na*	2851	2108	1994
S2	4467	na*	4782	2171	2235
S3	4654	na*	5255	2001	2144
S4	6417	na*	3494	1956	2052
R1	771	1479	823	973	953
R2	1563	2292	2042	1732	1672
R3	2512	4545	3324	2781	2854
R4	1504	4708	1493	1187	1157

Table 2b

Time Expressed as Percent of Time for the 4 Step Procedure

	4 Step	Convergence	Convergence Extrap.	Grouped Equal Int.	Grouped Centiles
S1	100	na*	82	61	57
S2	100	na*	107	49	50
S3	100	na*	113	43	46
S4	100	na*	54	30	32
R1	100	192	107	126	124
R2	100	147	131	111	107
R3	100	181	132	111	114
R4	100	313	99	79	77

na* - These were run on a different computer with a different convergence criterion.

Table 3

Summary Statistics Comparing Estimated Item Parameters to True Parameters

Common Items in Form 1 and Form 2 for tests R2 and R3

	Correlation			Bias			RMSE		
	a	b	c	a	b	c	a	b	c
R2									
F1 No prior on c	0.963	0.992	0.680	0.022	0.040	0.019	0.086	0.128	0.056
F1 Prior on c	0.961	0.996	0.850	0.004	0.019	0.006	0.086	0.093	0.037
F2 No prior on c	0.928	0.996	0.824	0.013	0.013	0.009	0.125	0.085	0.043
F2 Prior on c	0.921	0.997	0.869	-0.002	0.004	0.002	0.128	0.079	0.034
F2 No prior on c, Fix Comc	0.929	0.997	0.854	0.012	0.008	0.006	0.124	0.079	0.039
F2 Prior on c, Fix Mean	0.921	0.997	0.867	-0.002	0.005	0.003	0.128	0.079	0.034
R3									
F1 No prior on c	0.934	0.990	0.749	-0.006	0.007	0.006	0.113	0.191	0.054
F1 Prior on c	0.940	0.992	0.723	-0.019	0.005	0.003	0.111	0.165	0.051
F2 No prior on c	0.969	0.991	0.874	-0.006	-0.014	-0.005	0.096	0.215	0.044
F2 Prior on c	0.965	0.993	0.818	-0.025	-0.029	-0.019	0.098	0.224	0.047
F2 No prior on c, Fix Comc	0.969	0.991	0.875	-0.007	-0.018	-0.007	0.096	0.217	0.044
F2 Prior on c, Fix Mean	0.968	0.993	0.849	-0.009	-0.006	-0.004	0.091	0.209	0.039

Table 4

Summary Statistics Comparing Form 2 Item Parameters to Form 1 Item Parameters
Common Items in Form 1 and Form 2 for tests R2 and R3, real Math and Verbal

	Correlation			Bias			RMSE		
	a	b	c	a	b	c	a	b	c
R2									
F2 No prior on c	0.894	0.993	0.704	-0.009	-0.026	-0.010	0.150	0.115	0.054
F2 Prior on c	0.887	0.996	0.817	-0.006	-0.015	-0.003	0.151	0.086	0.035
F2 No prior on c, Fix Comc	0.895	0.993	0.697	-0.010	-0.031	-0.013	0.149	0.119	0.055
F2 Prior on c, Fix Mean	0.887	0.996	0.817	-0.006	-0.013	-0.003	0.150	0.085	0.035
R3									
F2 No prior on c	0.886	0.980	0.659	0.000	-0.021	-0.011	0.172	0.293	0.070
F2 Prior on c	0.883	0.984	0.568	-0.007	-0.034	-0.021	0.167	0.282	0.062
F2 No prior on c, Fix Comc	0.886	0.979	0.657	-0.001	-0.025	-0.013	0.172	0.298	0.071
F2 Prior on c, Fix Mean	0.887	0.985	0.631	0.010	-0.012	-0.006	0.165	0.262	0.054
Math									
F2 No prior on c	0.985	0.979	0.880	0.062	-0.016	-0.001	0.120	0.202	0.030
F2 Prior on c	0.984	0.980	0.882	0.064	-0.006	0.005	0.124	0.193	0.026
F2 No prior on c, Fix Comc	0.985	0.979	0.882	0.062	-0.012	0.001	0.120	0.199	0.029
F2 Prior on c, Fix Mean	0.984	0.980	0.882	0.064	-0.006	0.005	0.124	0.193	0.026
Verbal									
F2 No prior on c	0.886	0.984	0.669	-0.048	-0.070	-0.027	0.174	0.264	0.075
F2 Prior on c	0.889	0.989	0.696	-0.044	-0.069	-0.024	0.167	0.231	0.058
F2 No prior on c, Fix Comc	0.886	0.984	0.665	-0.049	-0.072	-0.028	0.174	0.265	0.076
F2 Prior on c, Fix Mean	0.888	0.988	0.724	-0.035	-0.059	-0.015	0.166	0.236	0.053

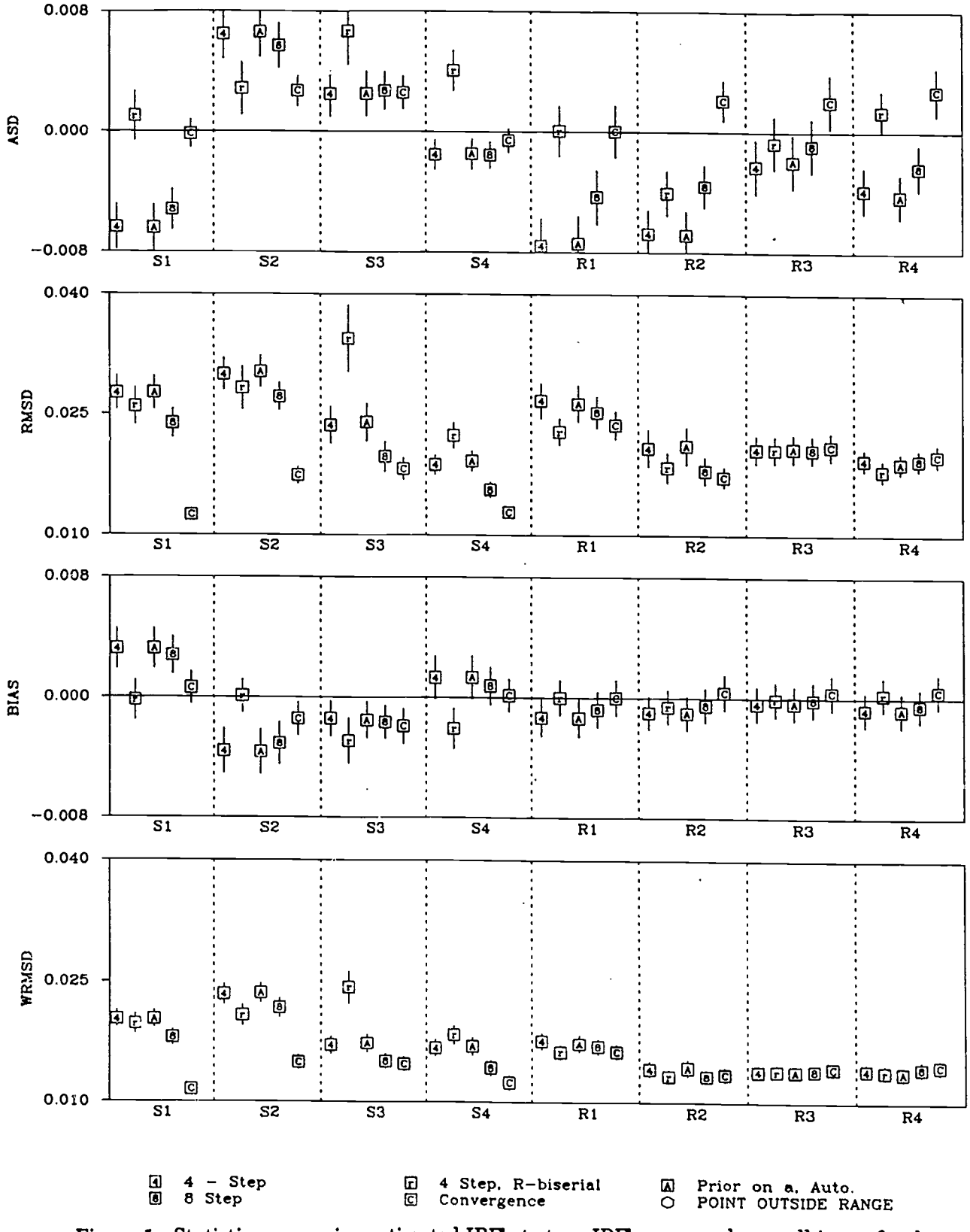

Figure 1. Statistics comparing estimated IRF's to true IRF's averaged over all items for the automatic methods and the run to convergence. The lines above and below the box plotted extend one standard error of the mean from the center of the box.

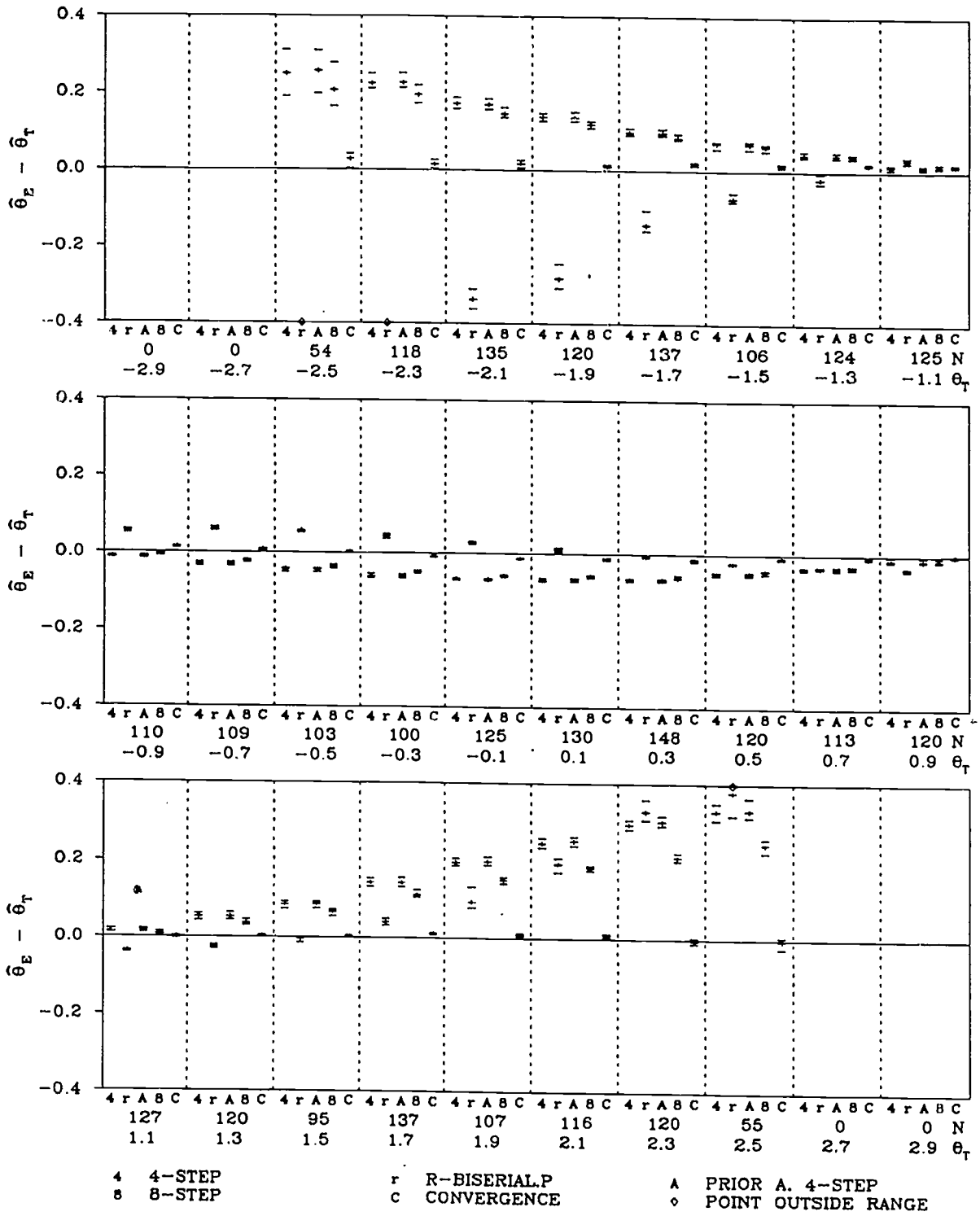

Figure 2. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Automatic methods and the run to convergence. Test S1.

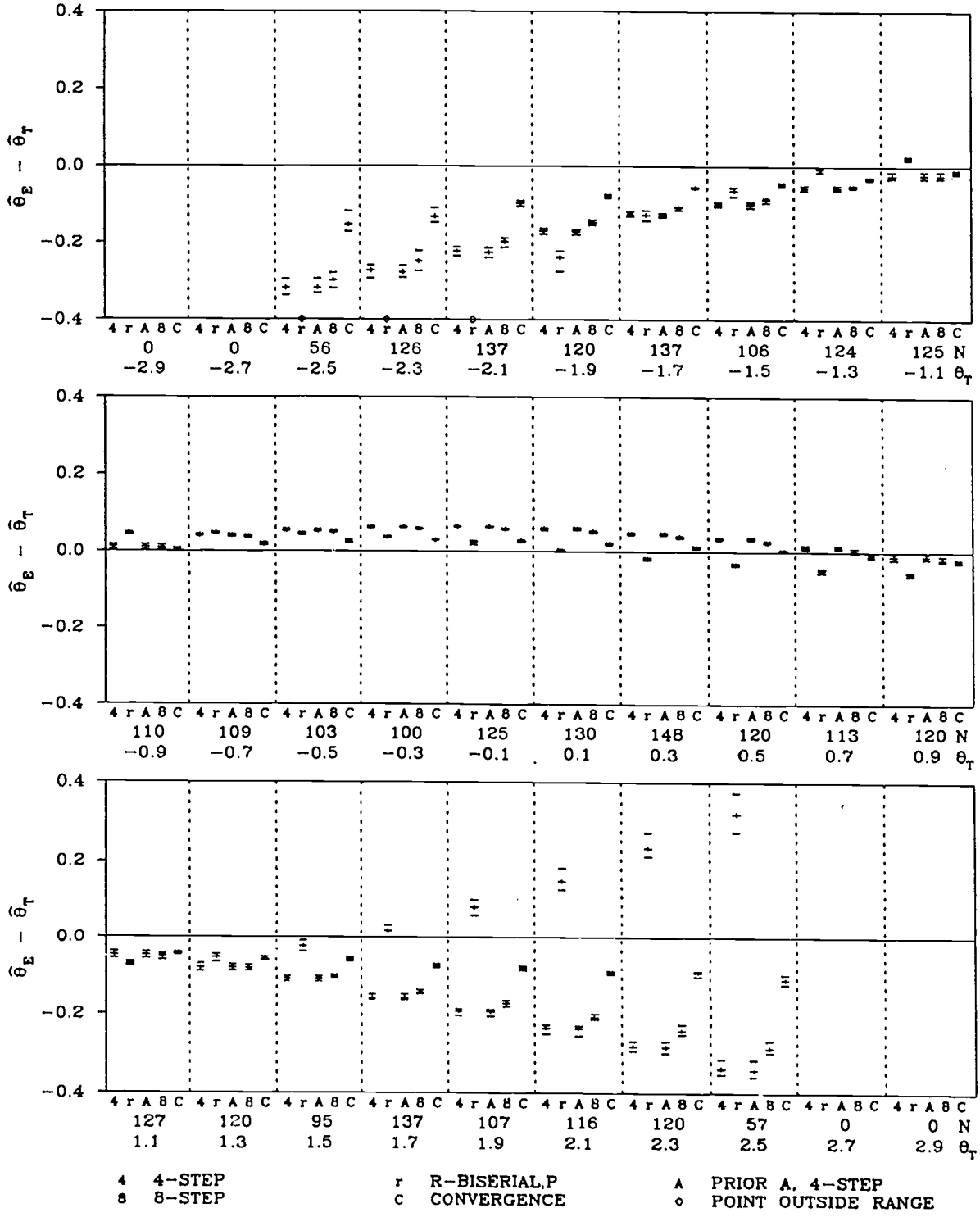

Figure 3. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Automatic methods and the run to convergence. Test S2.

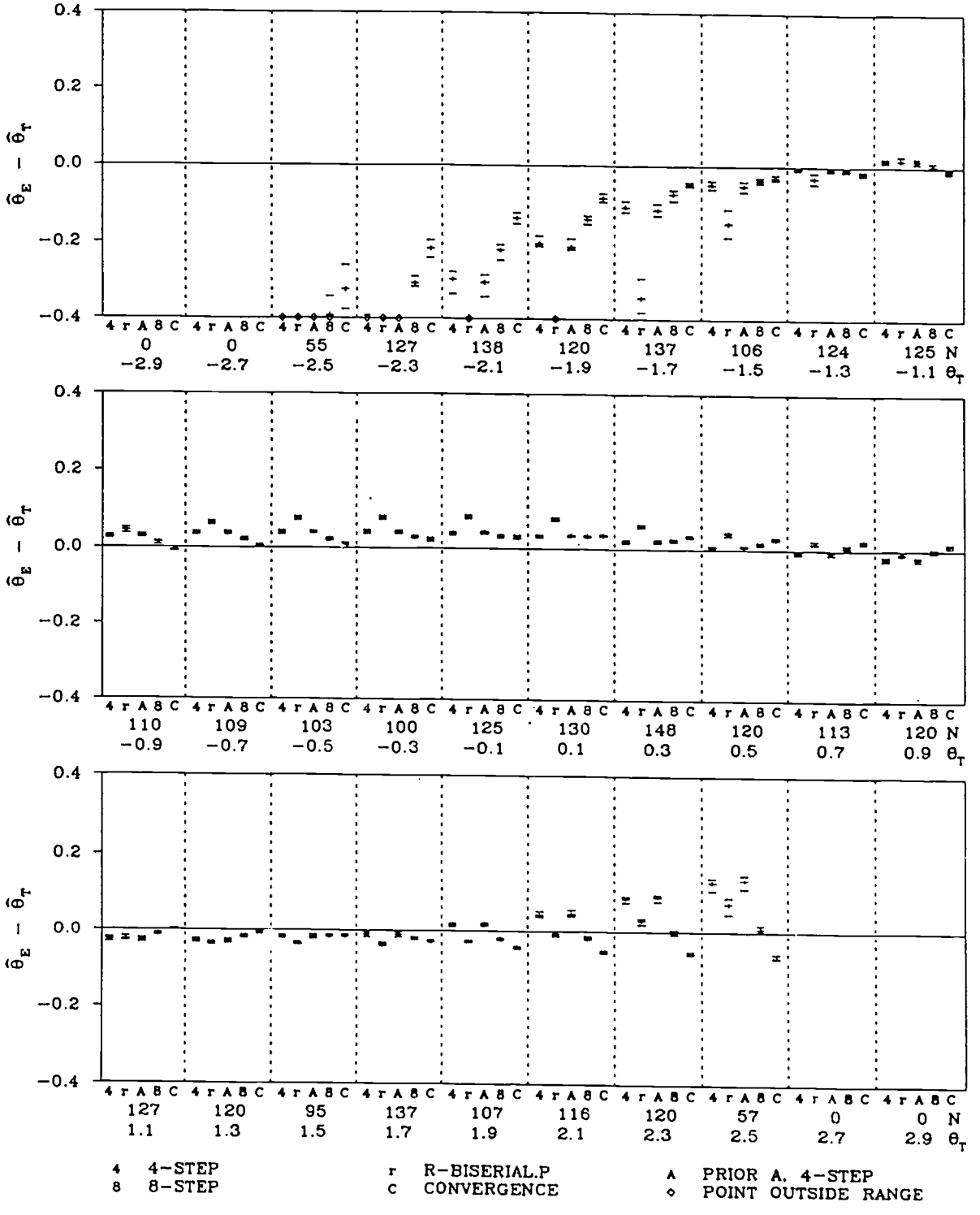

Figure 4. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Automatic methods and the run to convergence. Test S3.

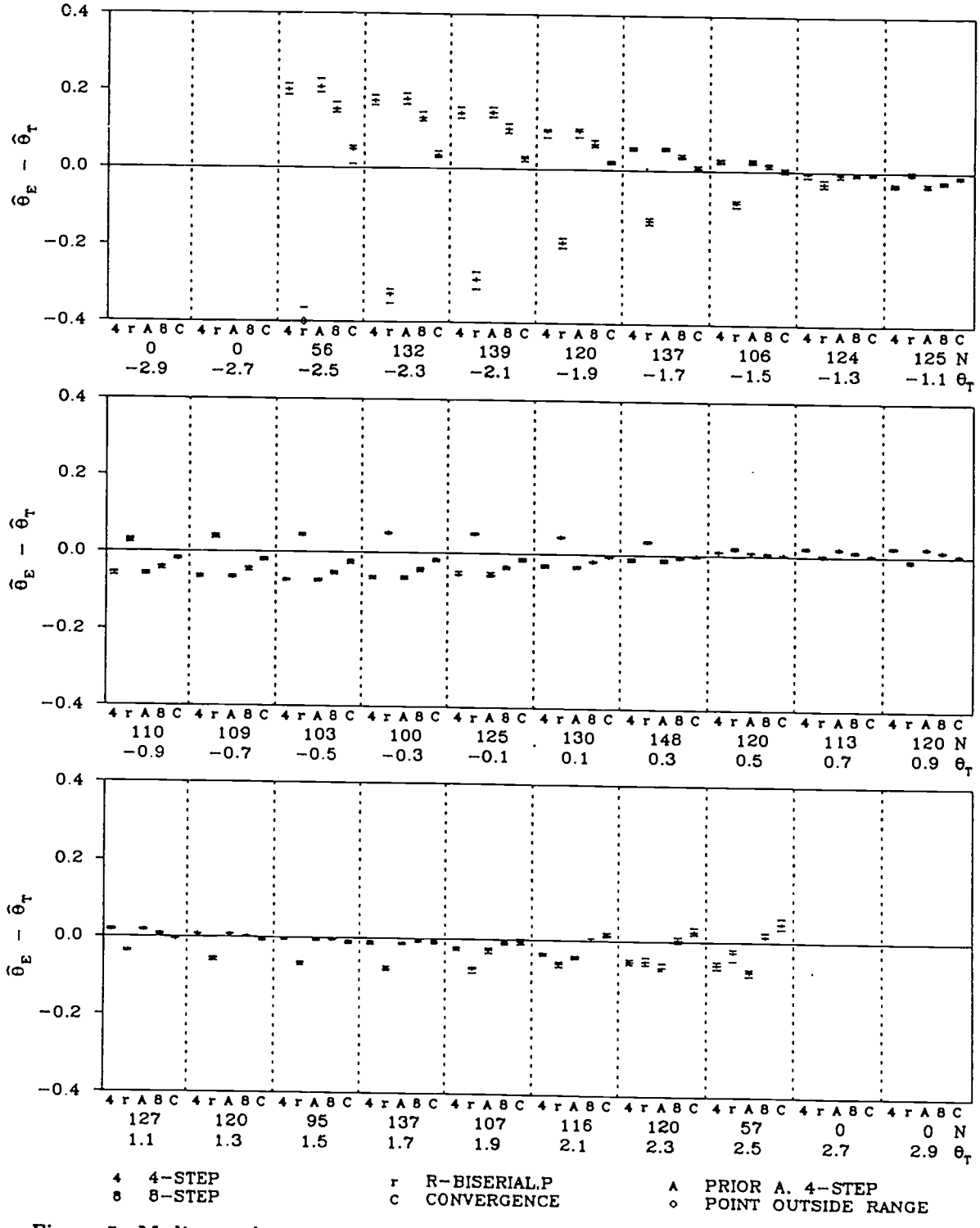

Figure 5. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Automatic methods and the run to convergence. Test S4.

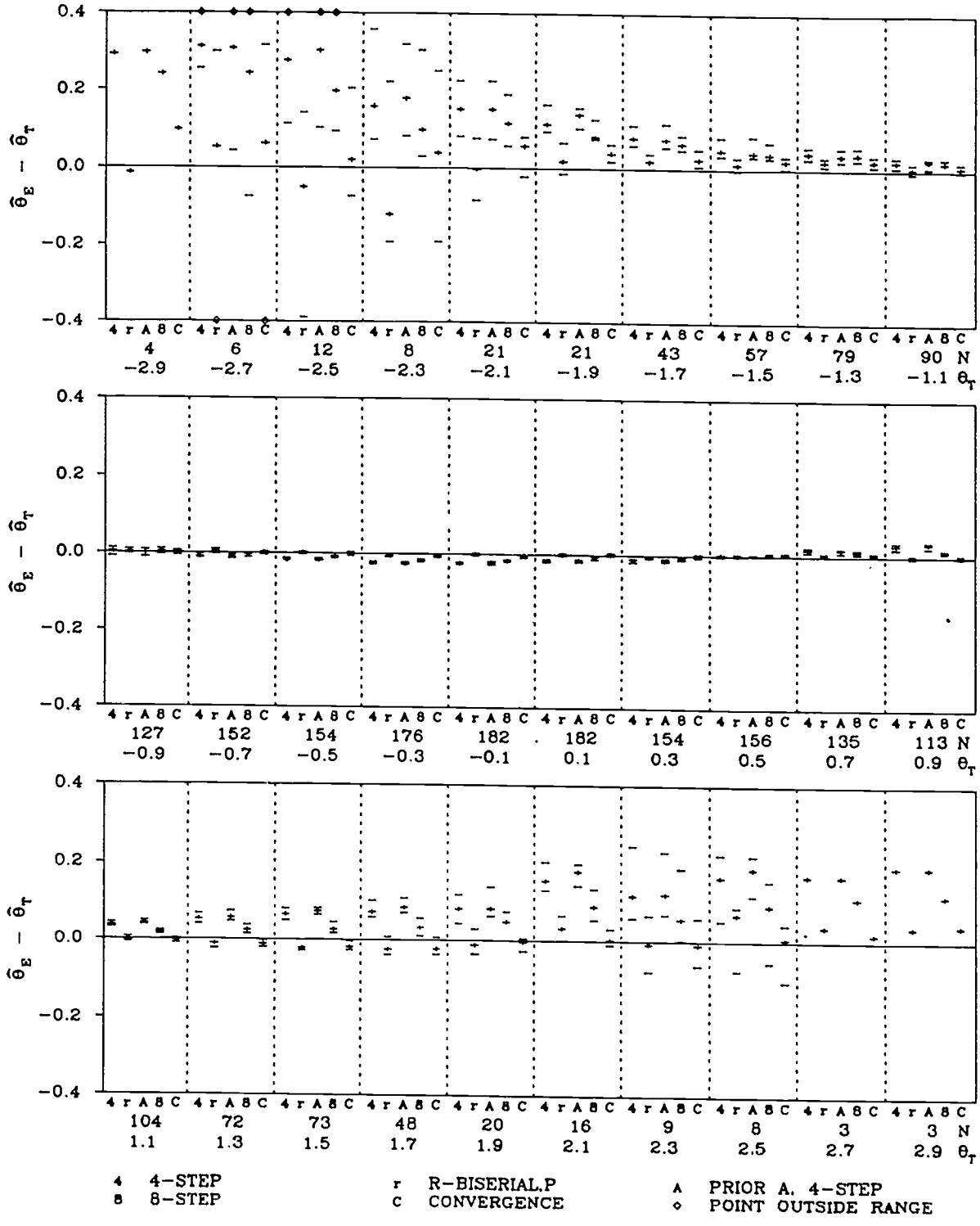

Figure 6. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Automatic methods and the run to convergence. Test R1.

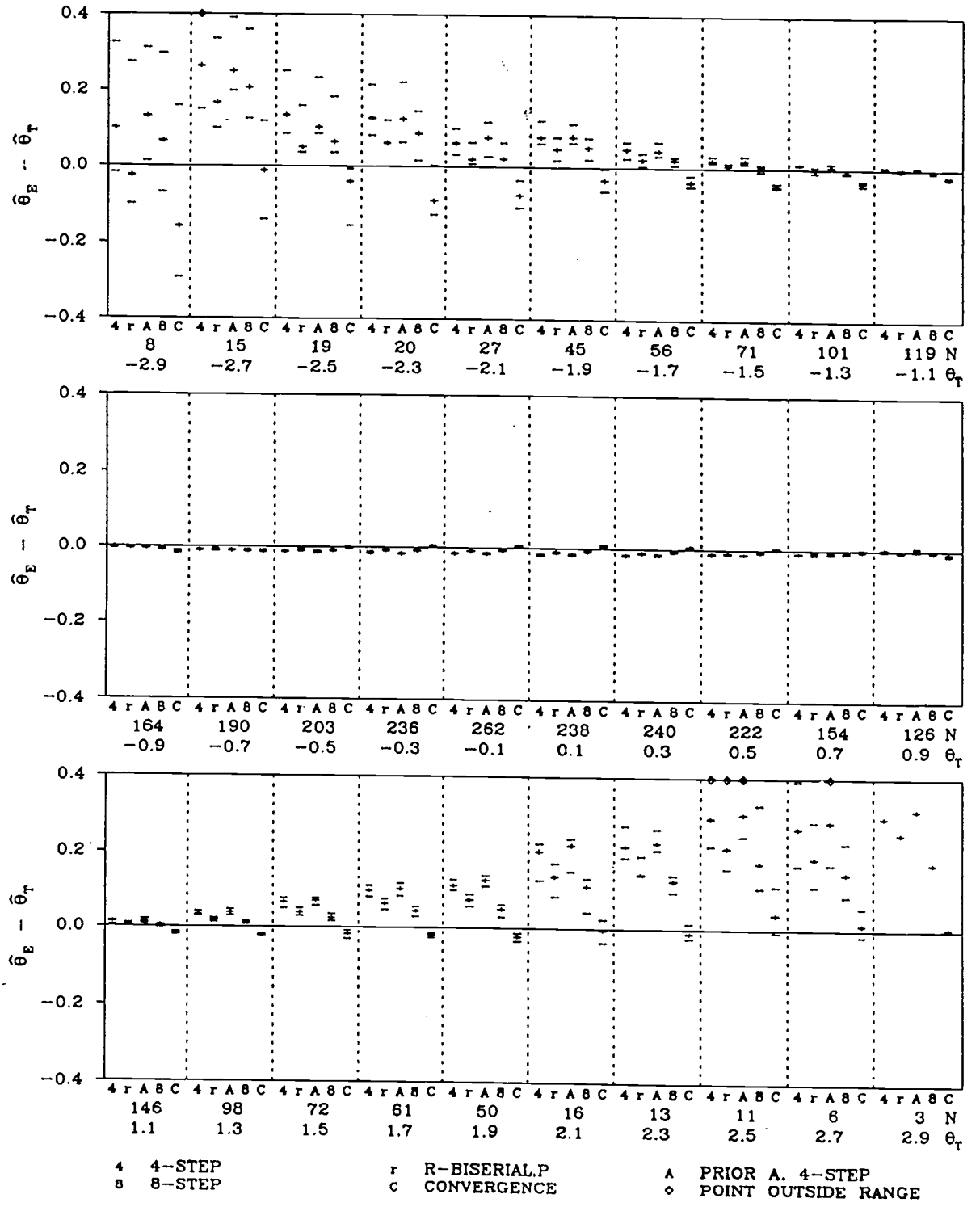

Figure 7. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Automatic methods and the run to convergence. Test R2.

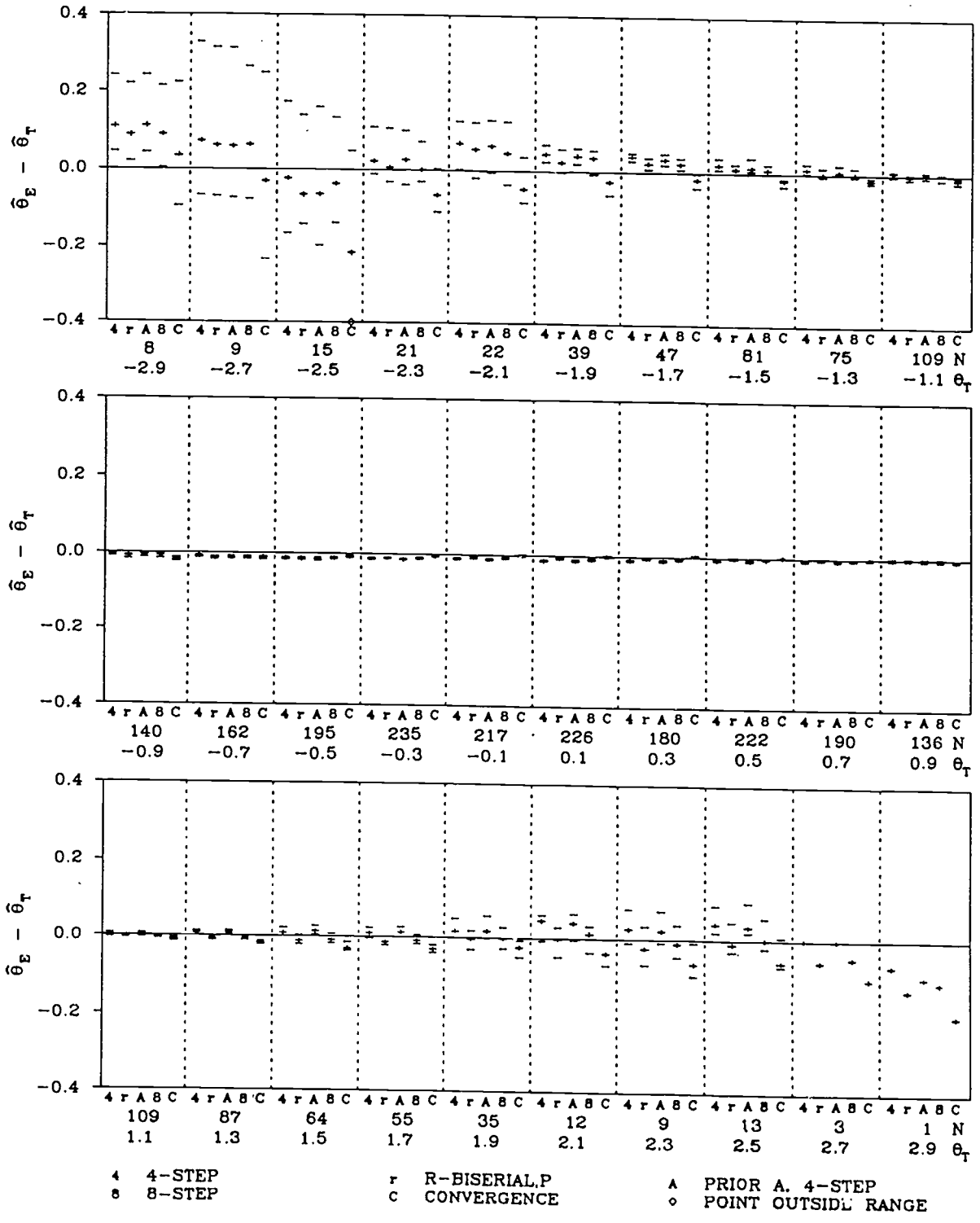

Figure 8. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Automatic methods and the run to convergence. Test R3.

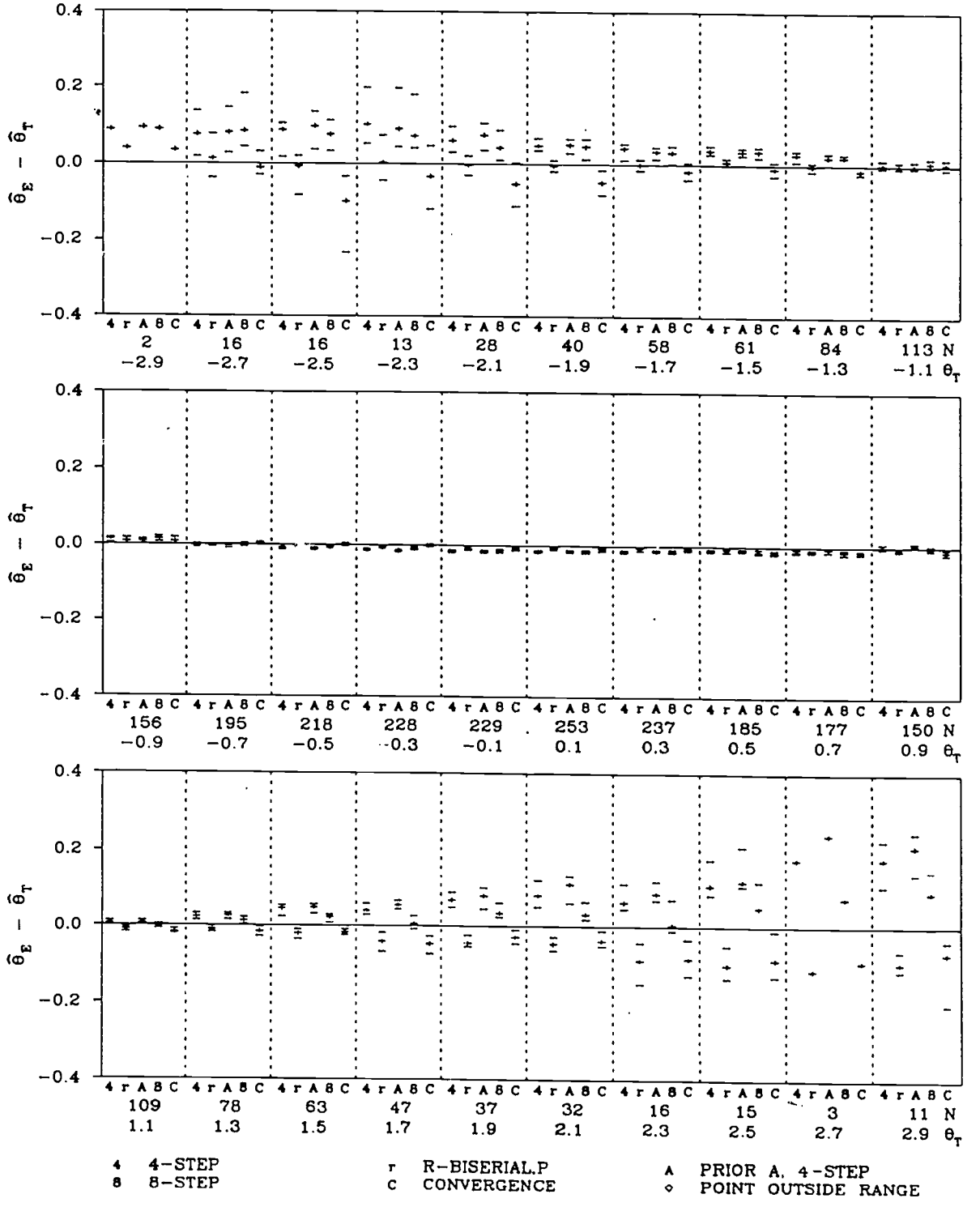

Figure 9. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Automatic methods and the run to convergence. Test R4.

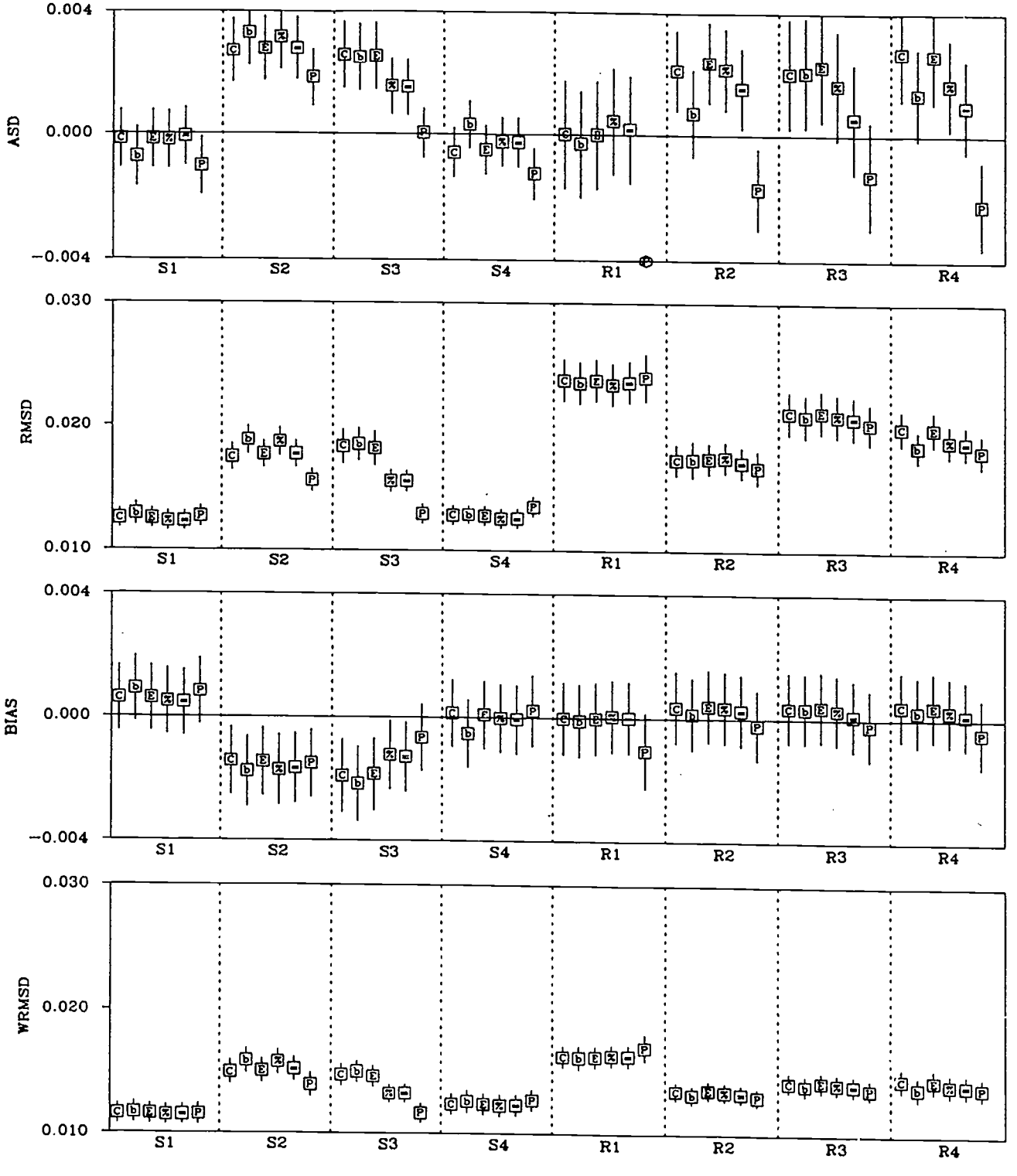

Figure 10. Statistics comparing estimated IRF's to true IRF's averaged over all items for the convergence methods. The lines above and below the box plotted extend one standard error of the mean from the center of the box.

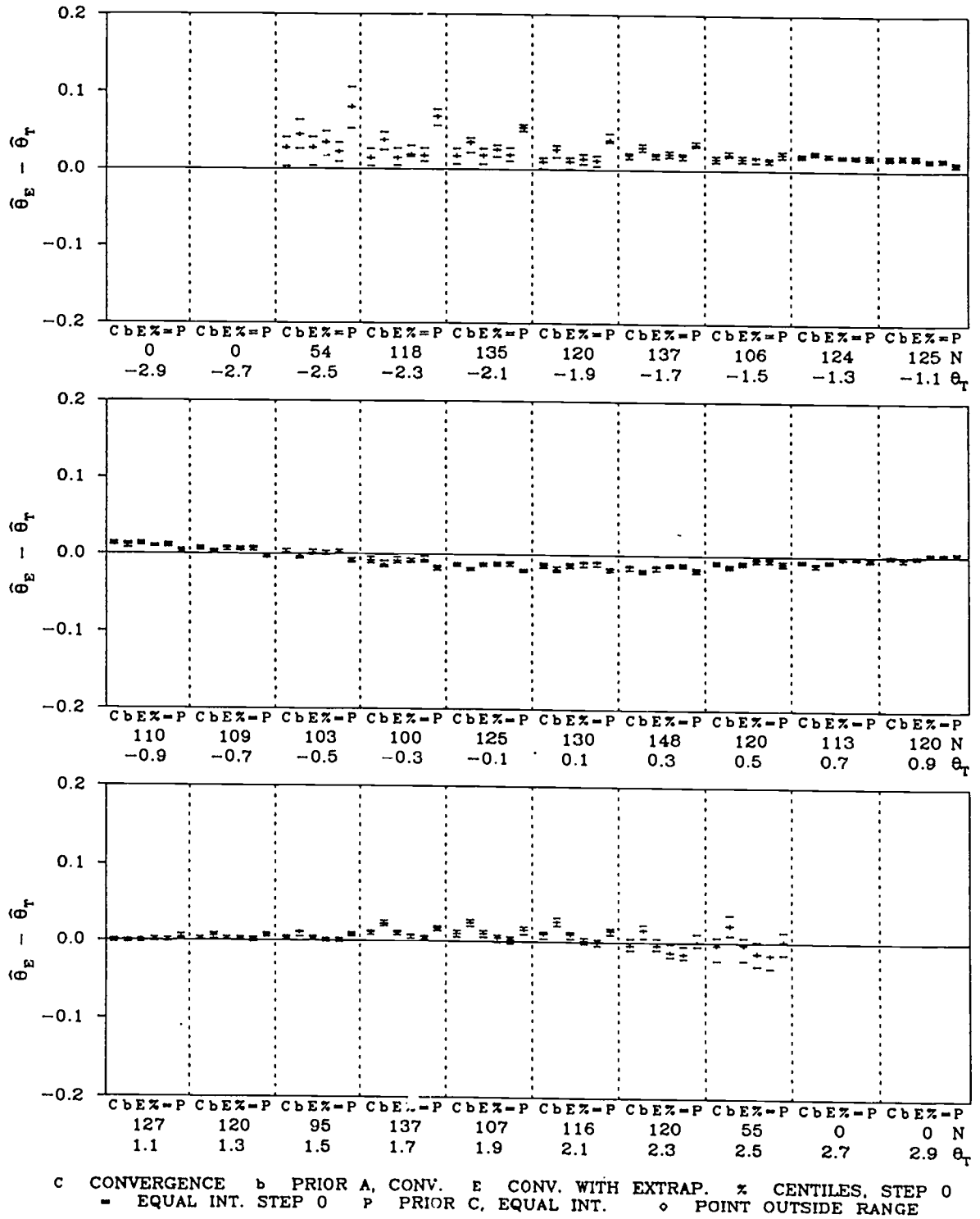

Figure 11. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Convergence methods. Test S1.

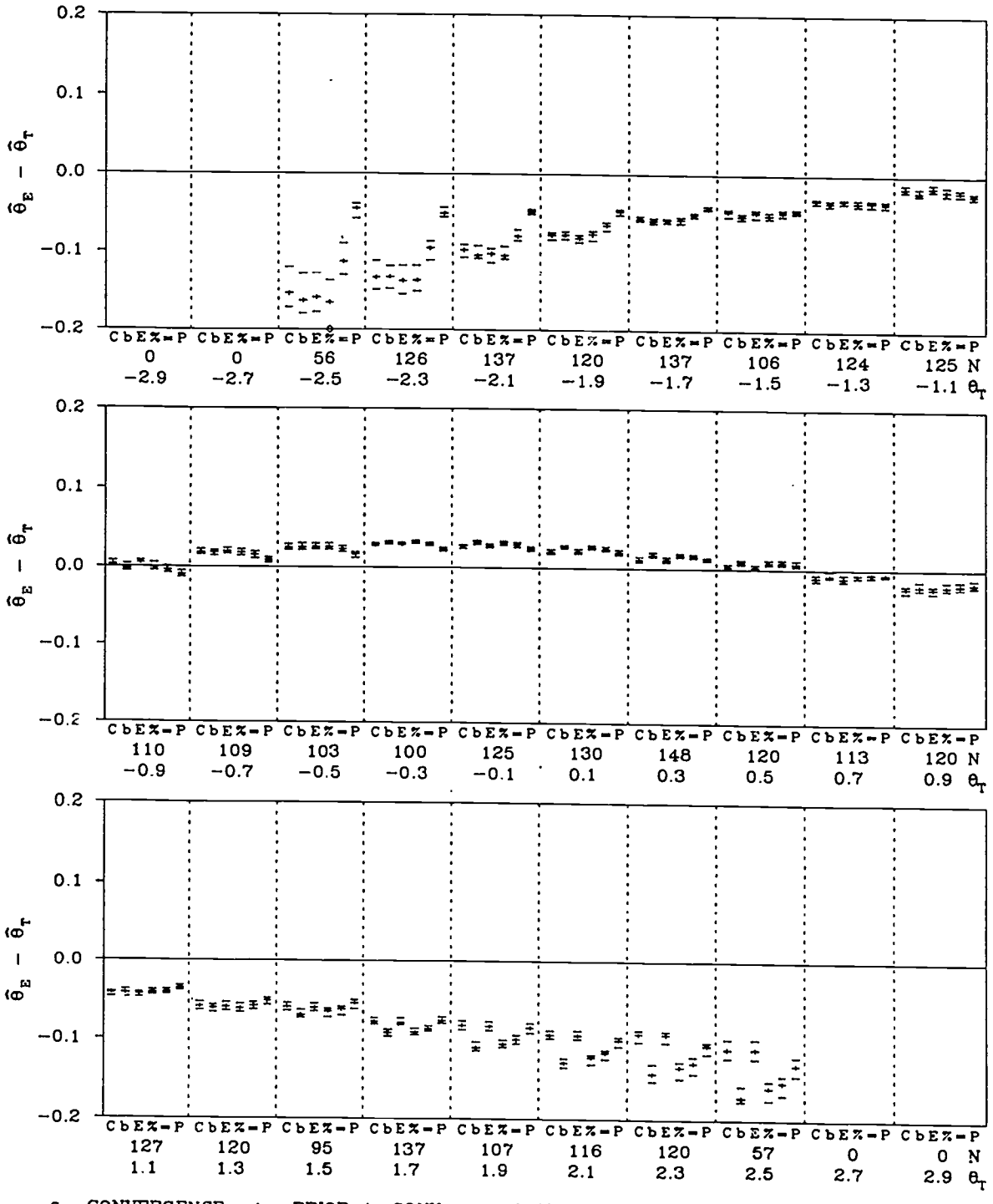

Figure 12. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Convergence methods. Test S2.

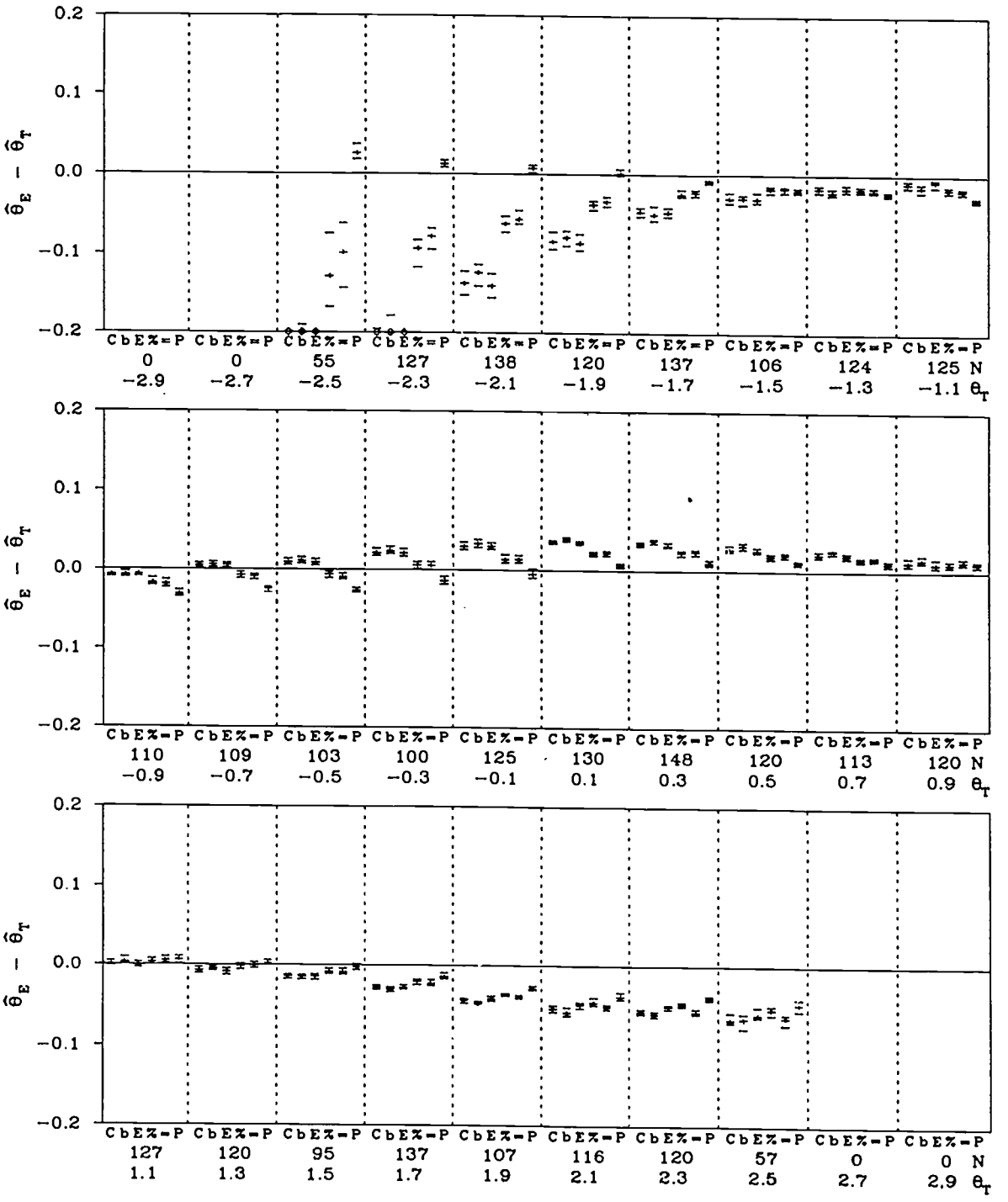

Figure 13. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Convergence methods. Test S3.

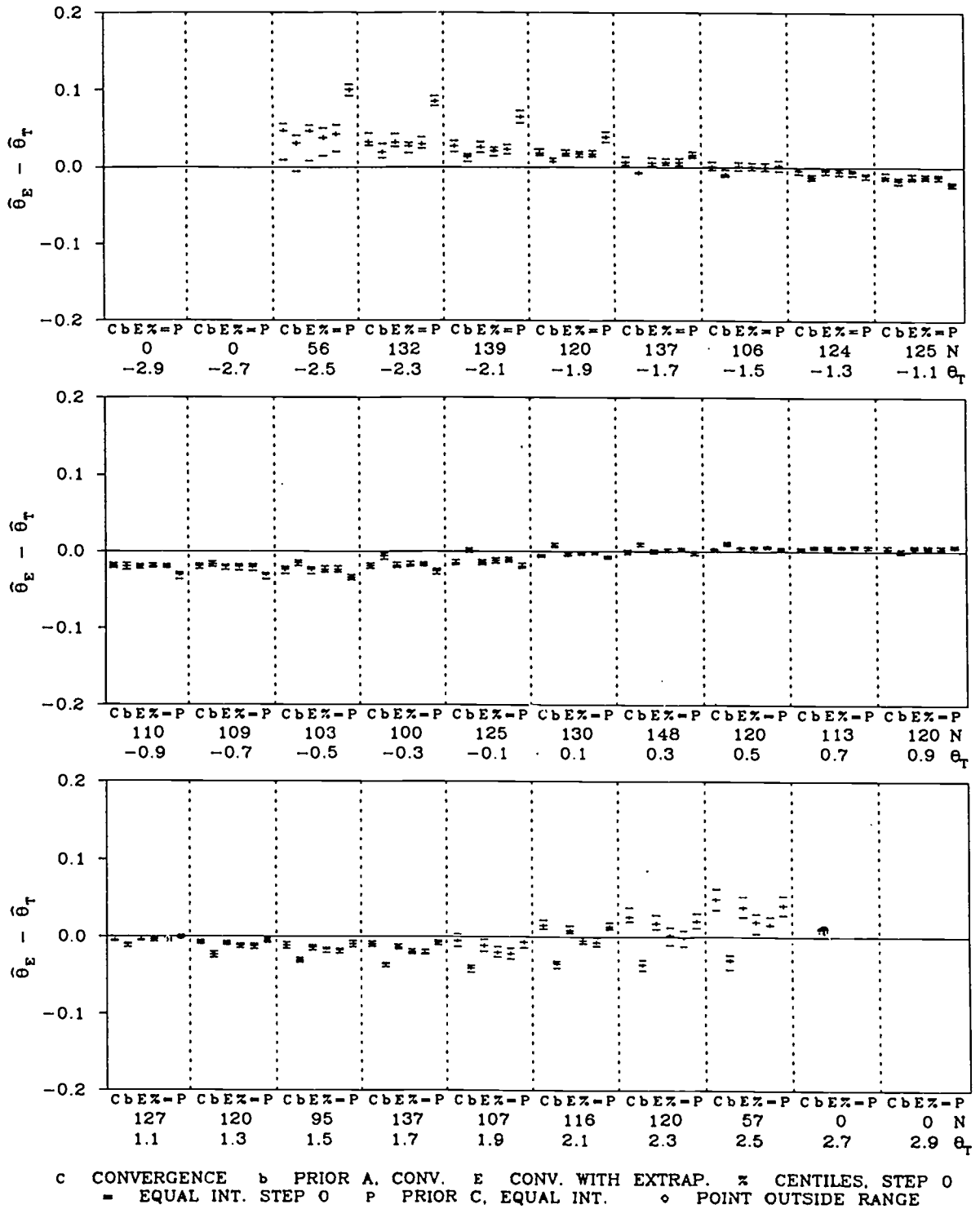

Figure 14. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Convergence methods. Test S4.

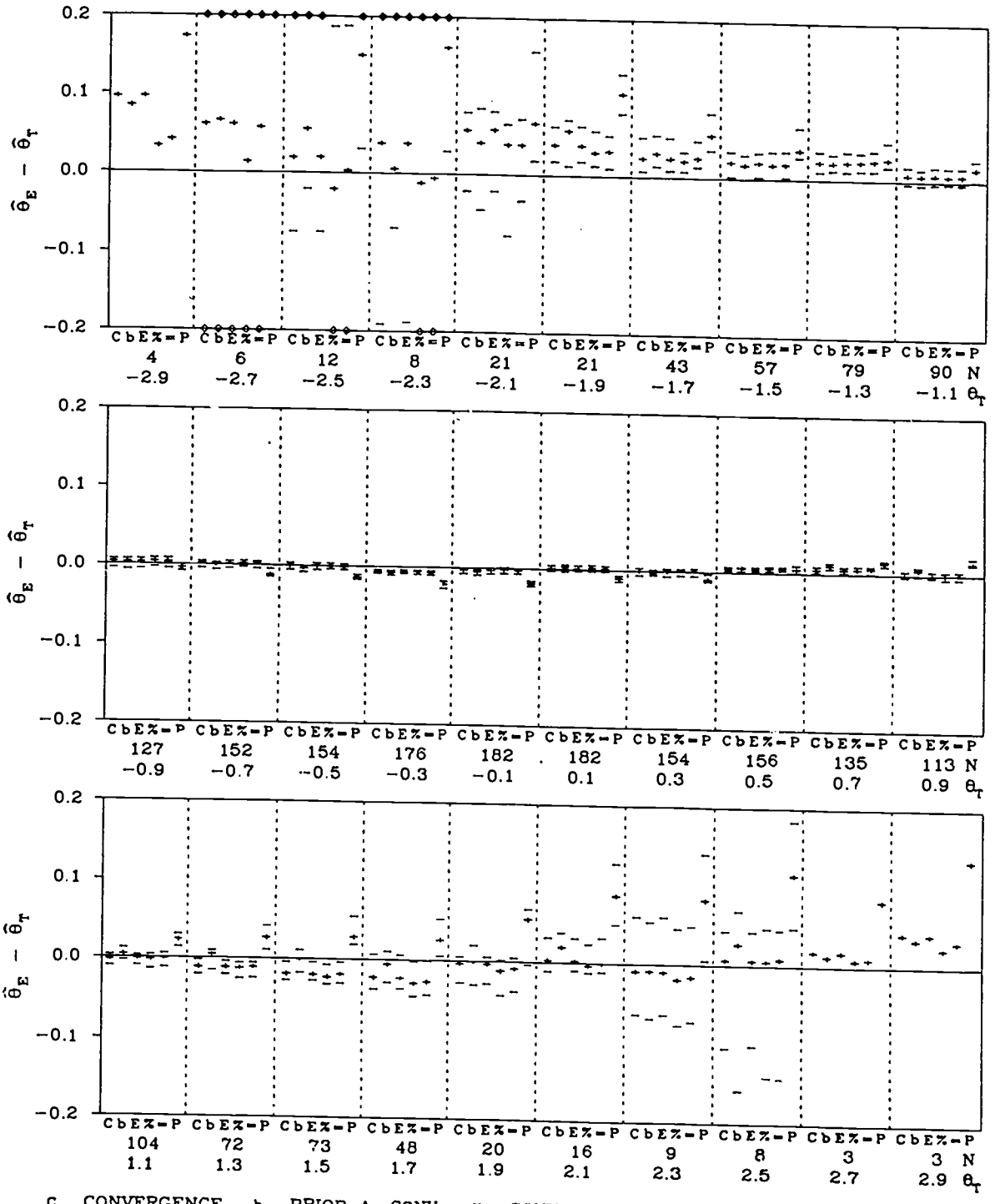

Figure 15. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Convergence methods. Test R1.

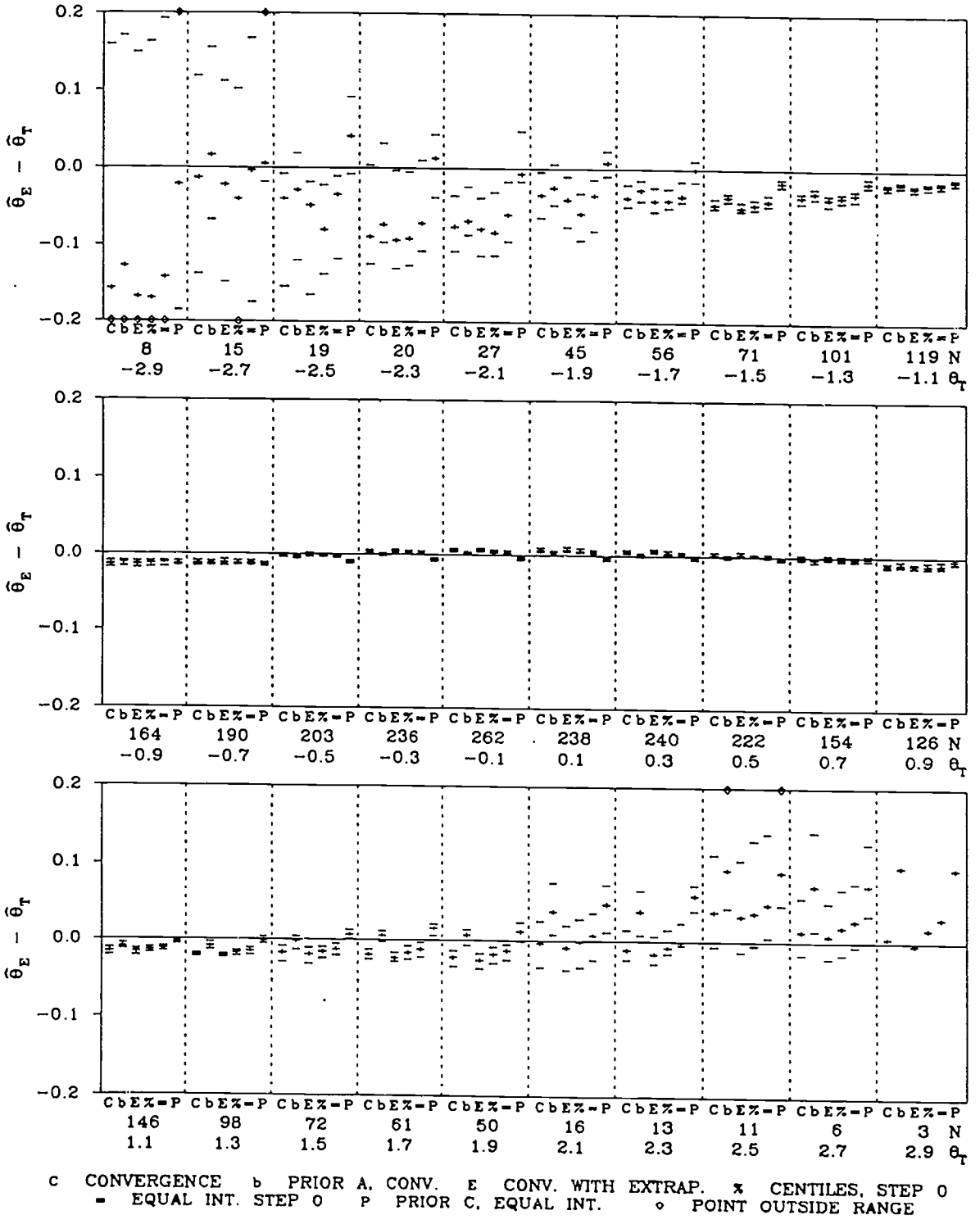

Figure 16. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Converg ce methods. Test R2.

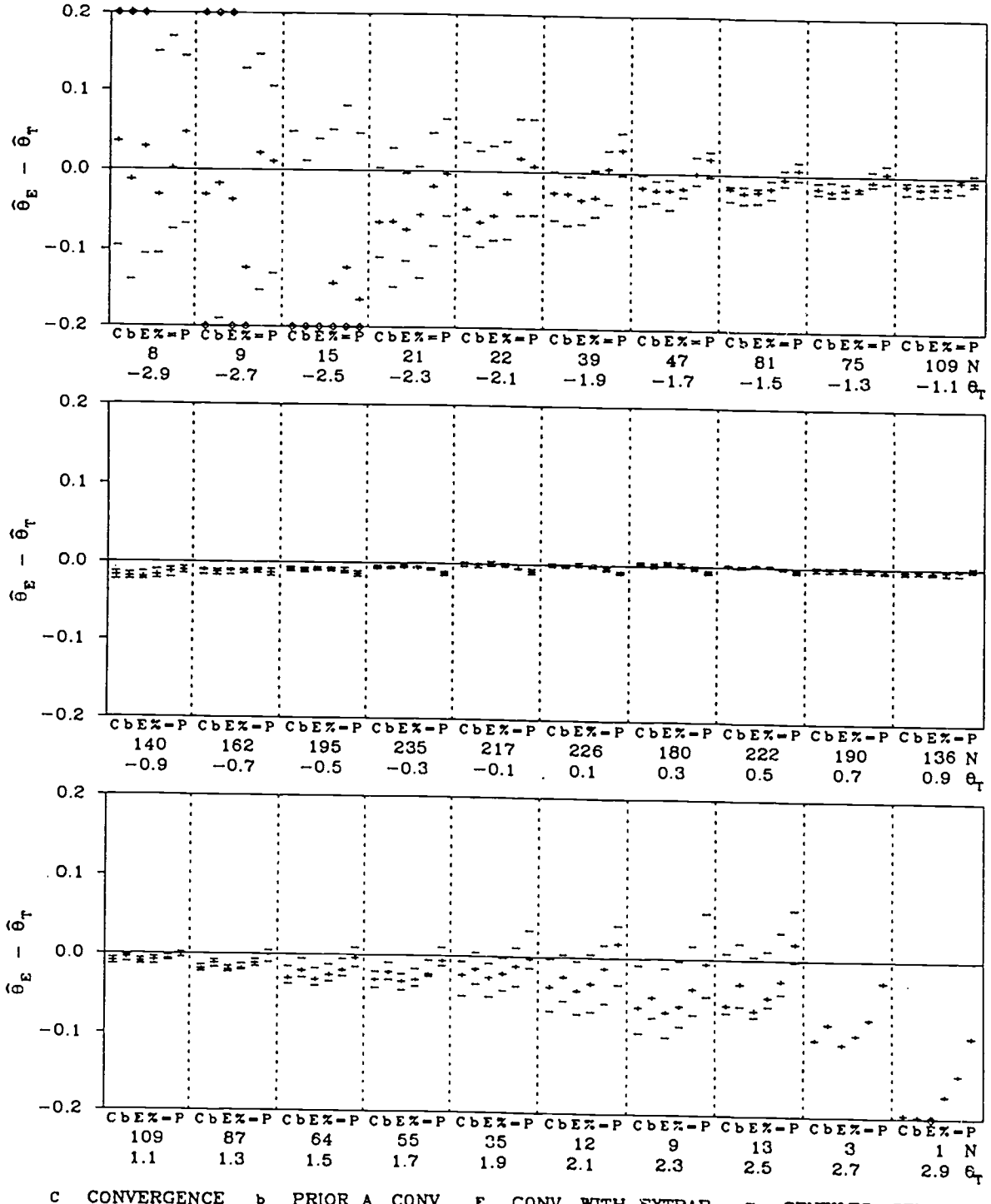

Figure 17. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Convergence methods. Test R3.

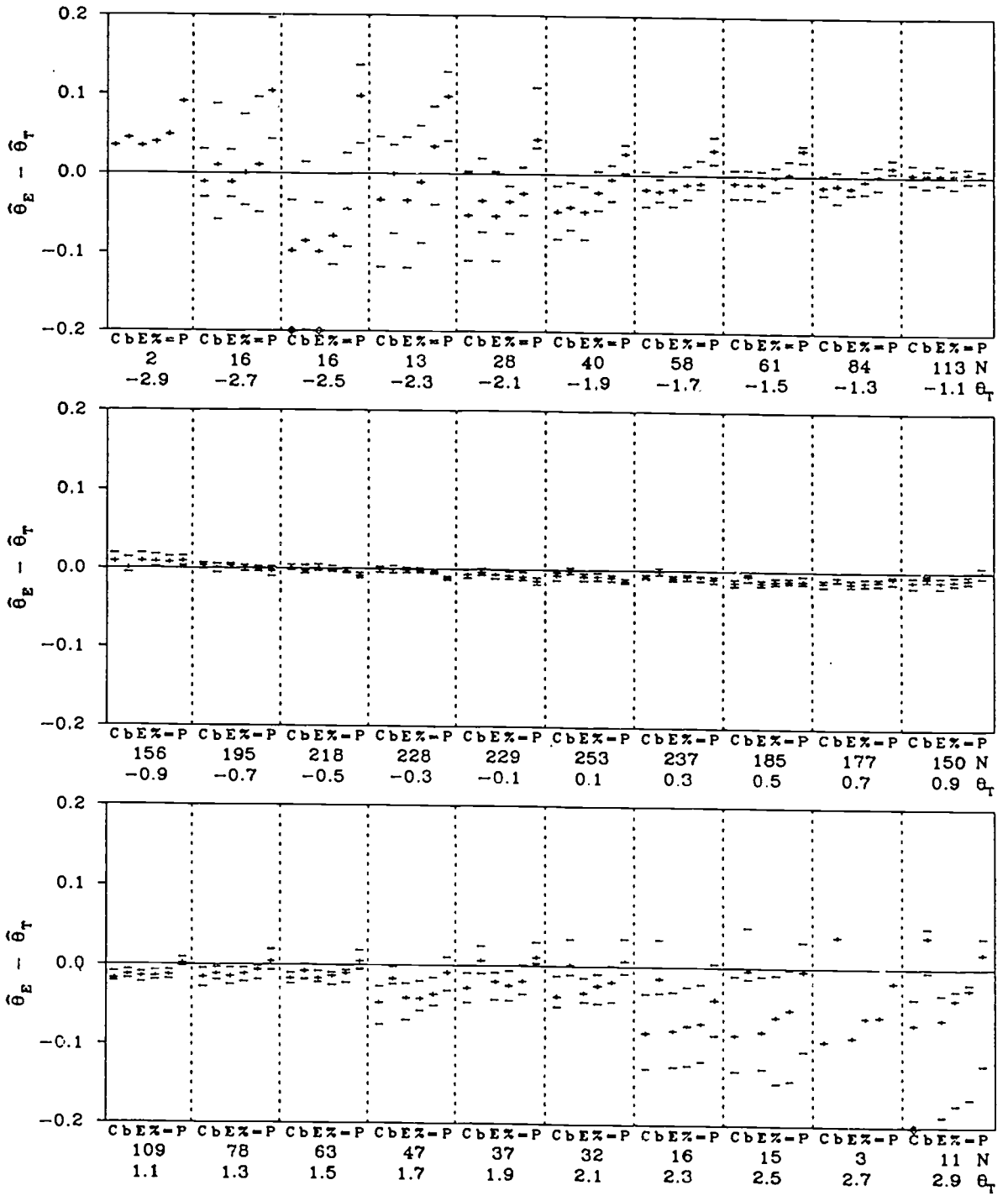

Figure 18. Median and 5% two-tailed confidence band of residuals between ability computed with estimated item parameters and ability computed with true item parameters. Median and confidence band conditional on true ability. Convergence methods. Test R4.

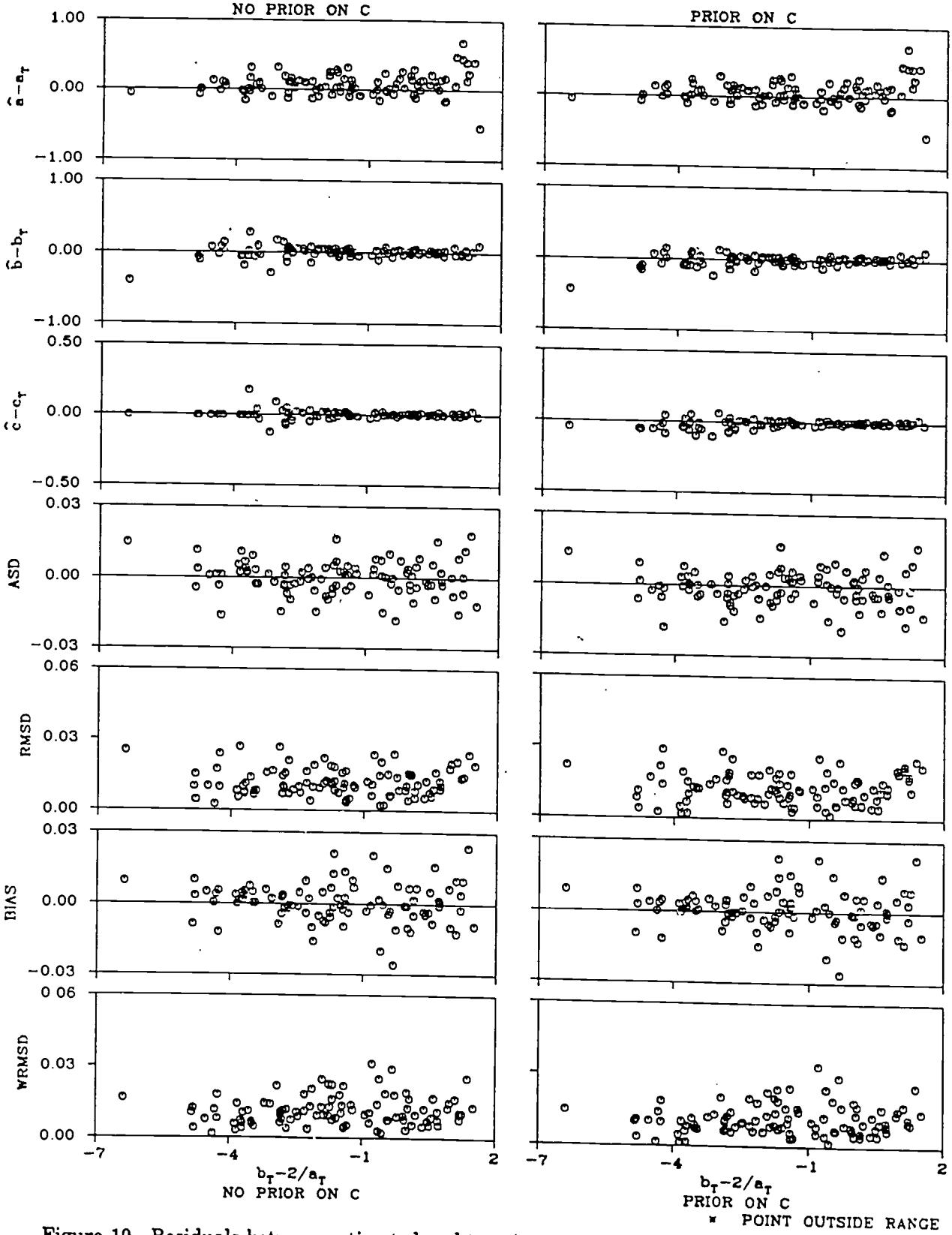

Figure 19. Residuals between estimated and true item parameters and statistics comparing estimated IRF's to true IRF's for the Method = with no prior on c and the Method P with a prior on c. All results plotted against the true b-2/a. Test S1.

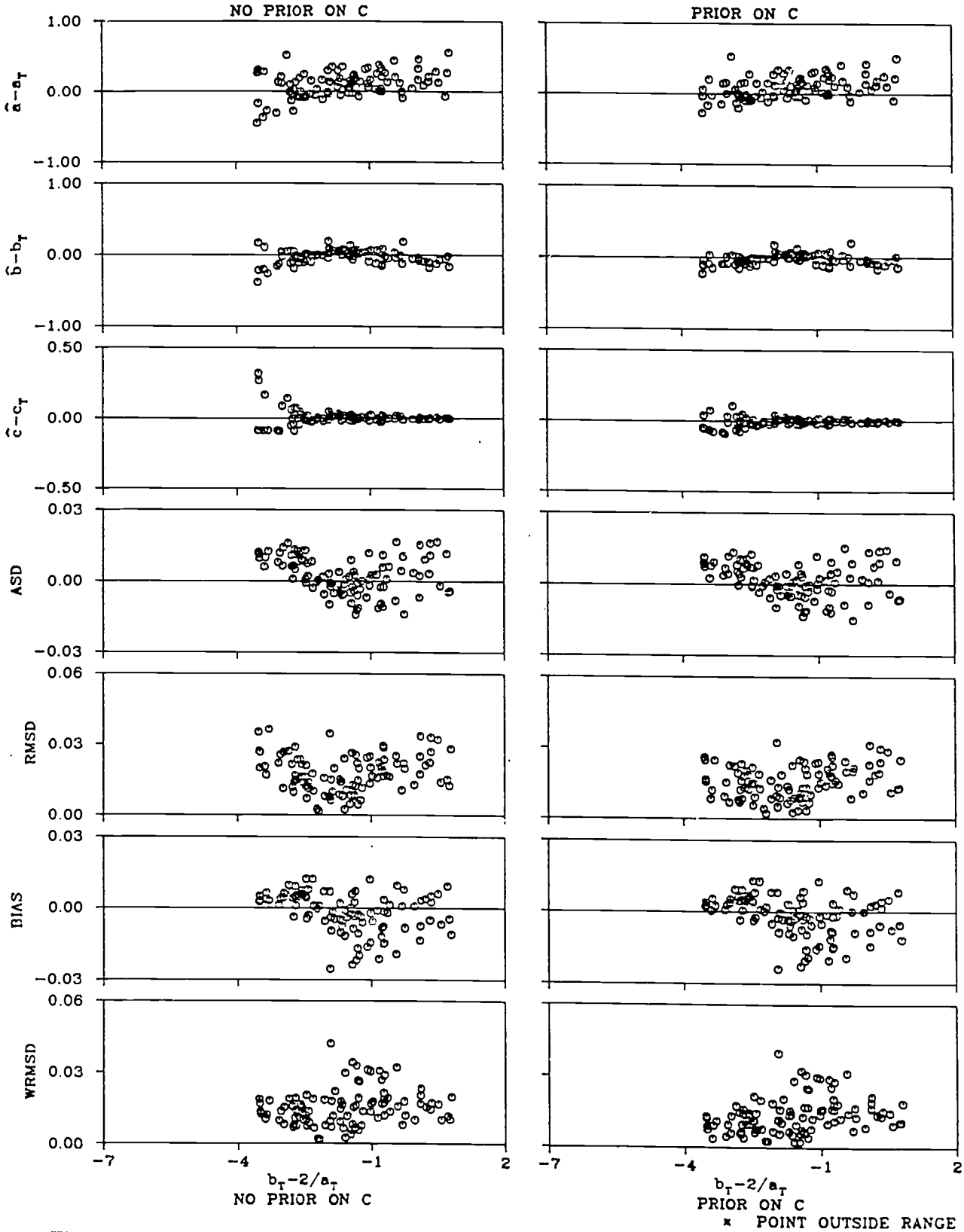

Figure 20. Residuals between estimated and true item parameters and statistics comparing estimated IRF's to true IRF's for the Method = with no prior on c and the Method P with a prior on c. All results plotted against the true b-2/a. Test S2.

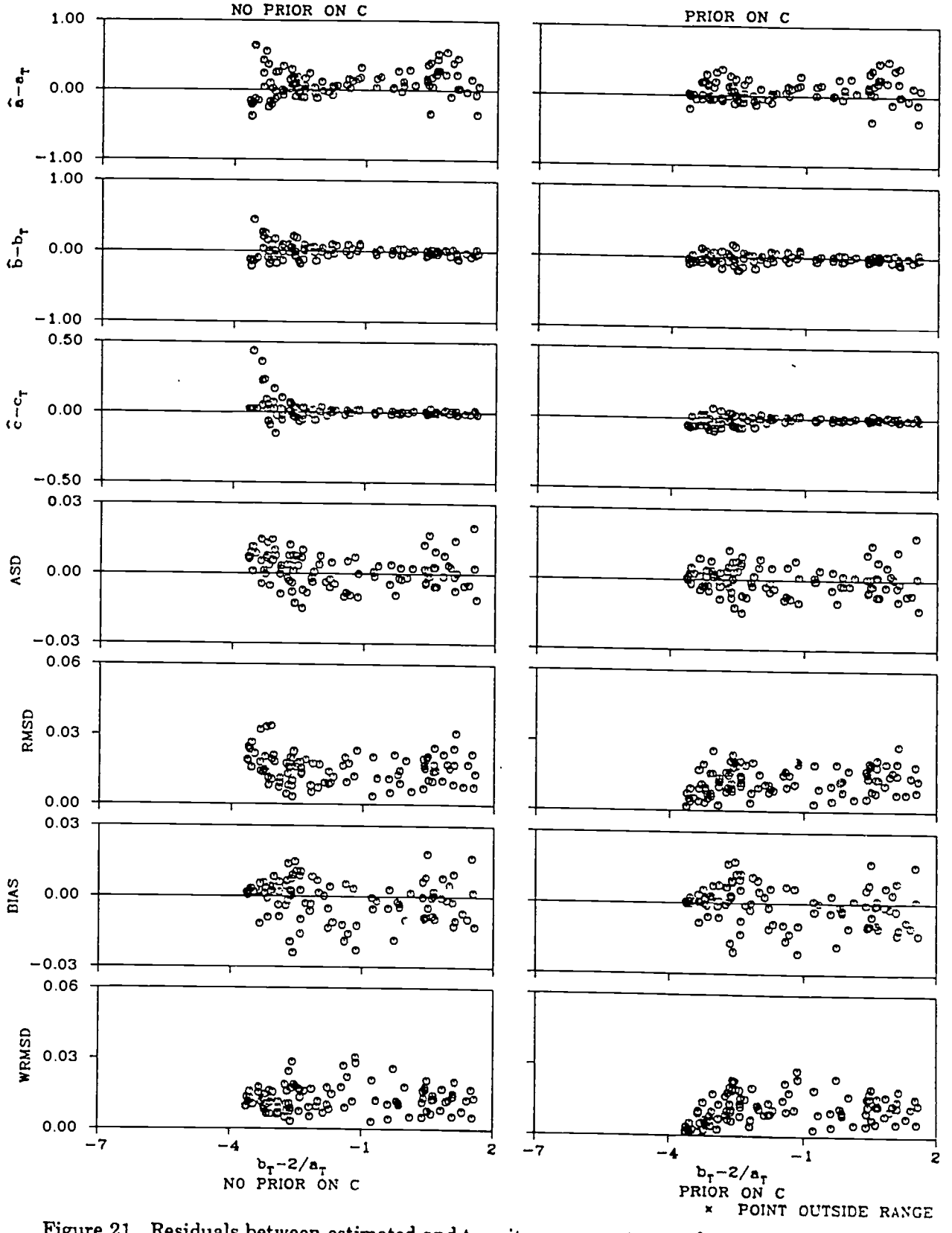

Figure 21. Residuals between estimated and true item parameters and statistics comparing estimated IRF's to true IRF's for the Method = with no prior on c and the Method P with a prior on c. All results plotted against the true b-2/a. Test S3.

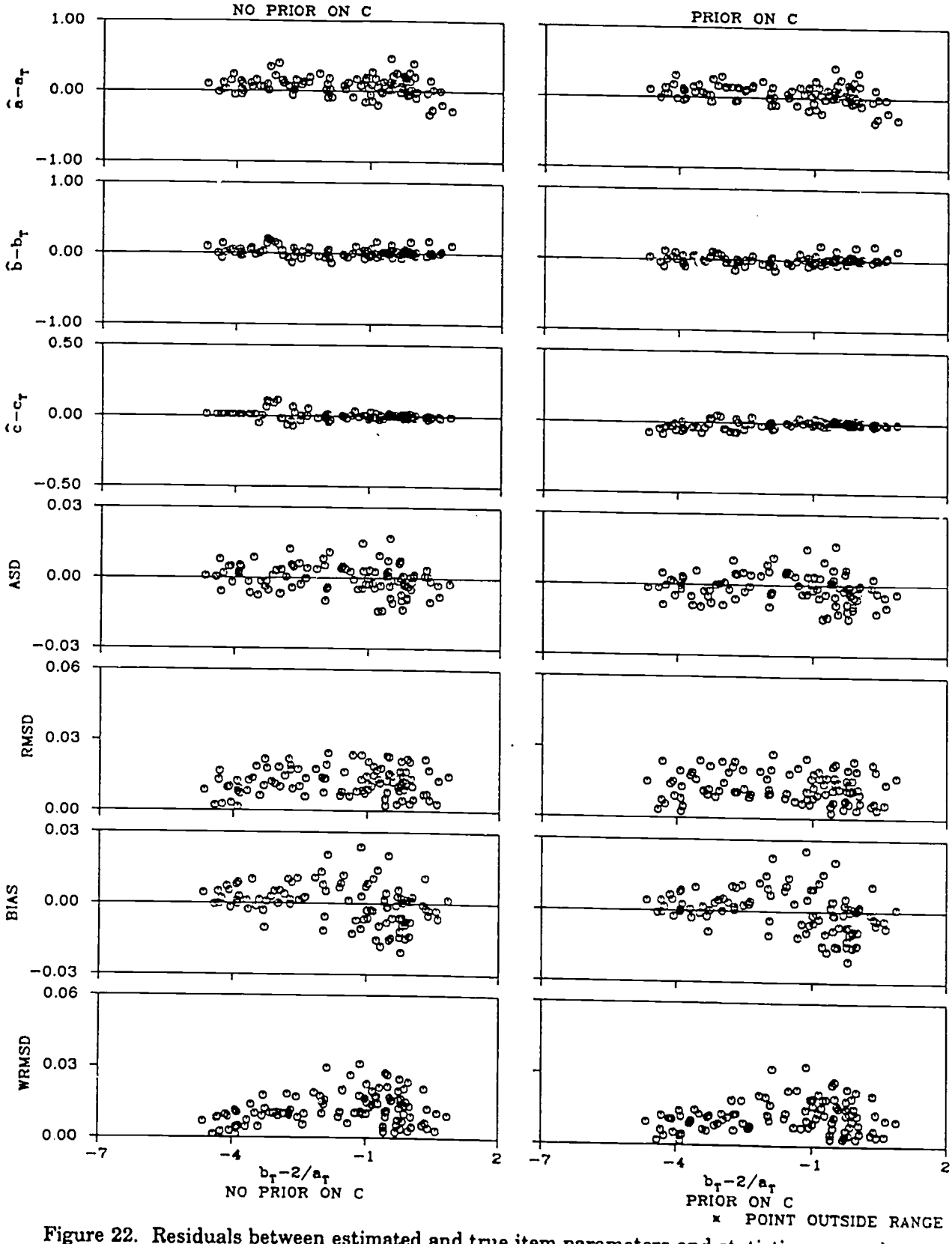

Figure 22. Residuals between estimated and true item parameters and statistics comparing estimated IRF's to true IRF's for the Method = with no prior on c and the Method P with a prior on c. All results plotted against the true b-2/a. Test S4.

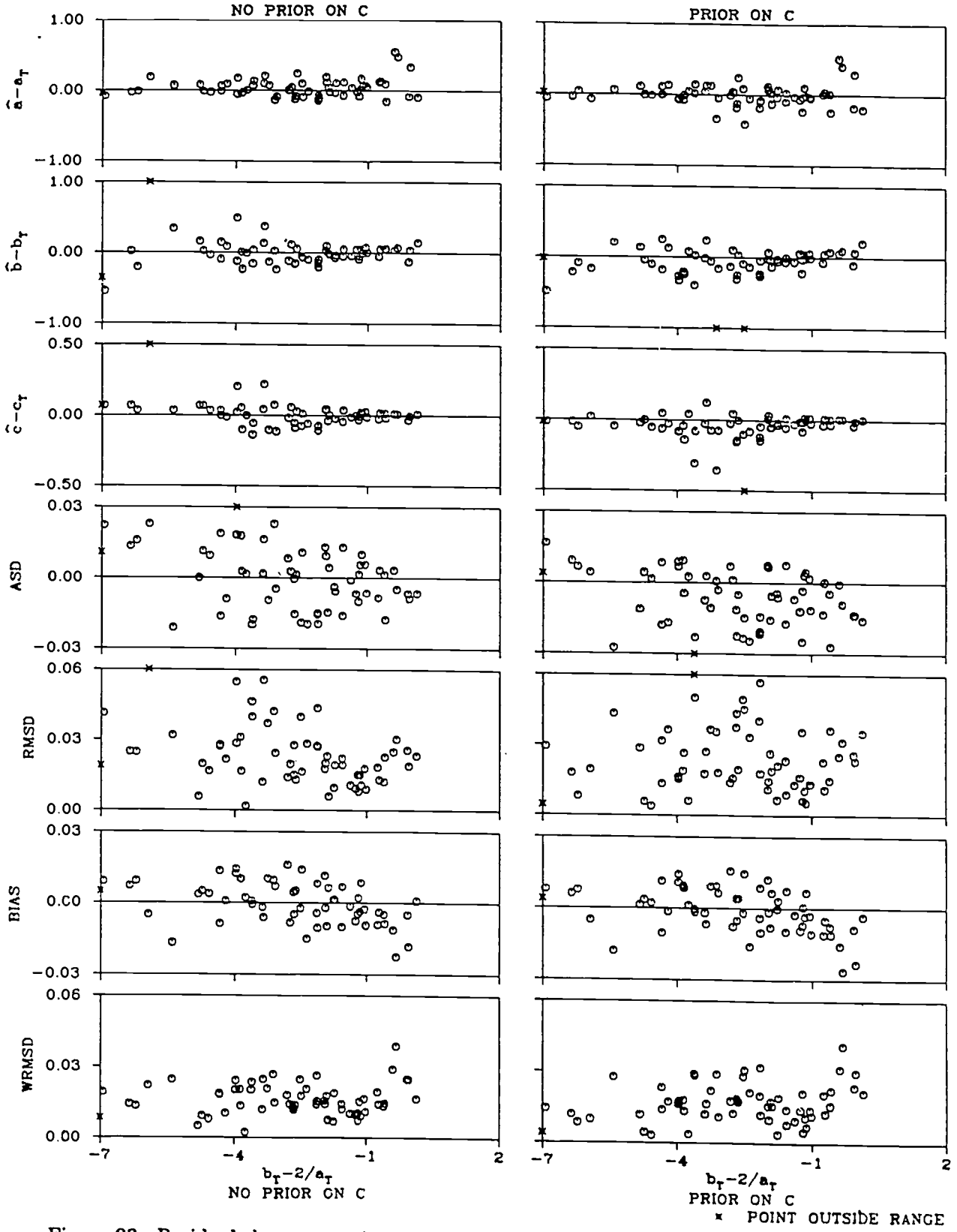

Figure 23. Residuals between estimated and true item parameters and statistics comparing estimated IRF's to true IRF's for the Method = with no prior on c and the Method P with a prior on c. All results plotted against the true b-2/a. Test R1.

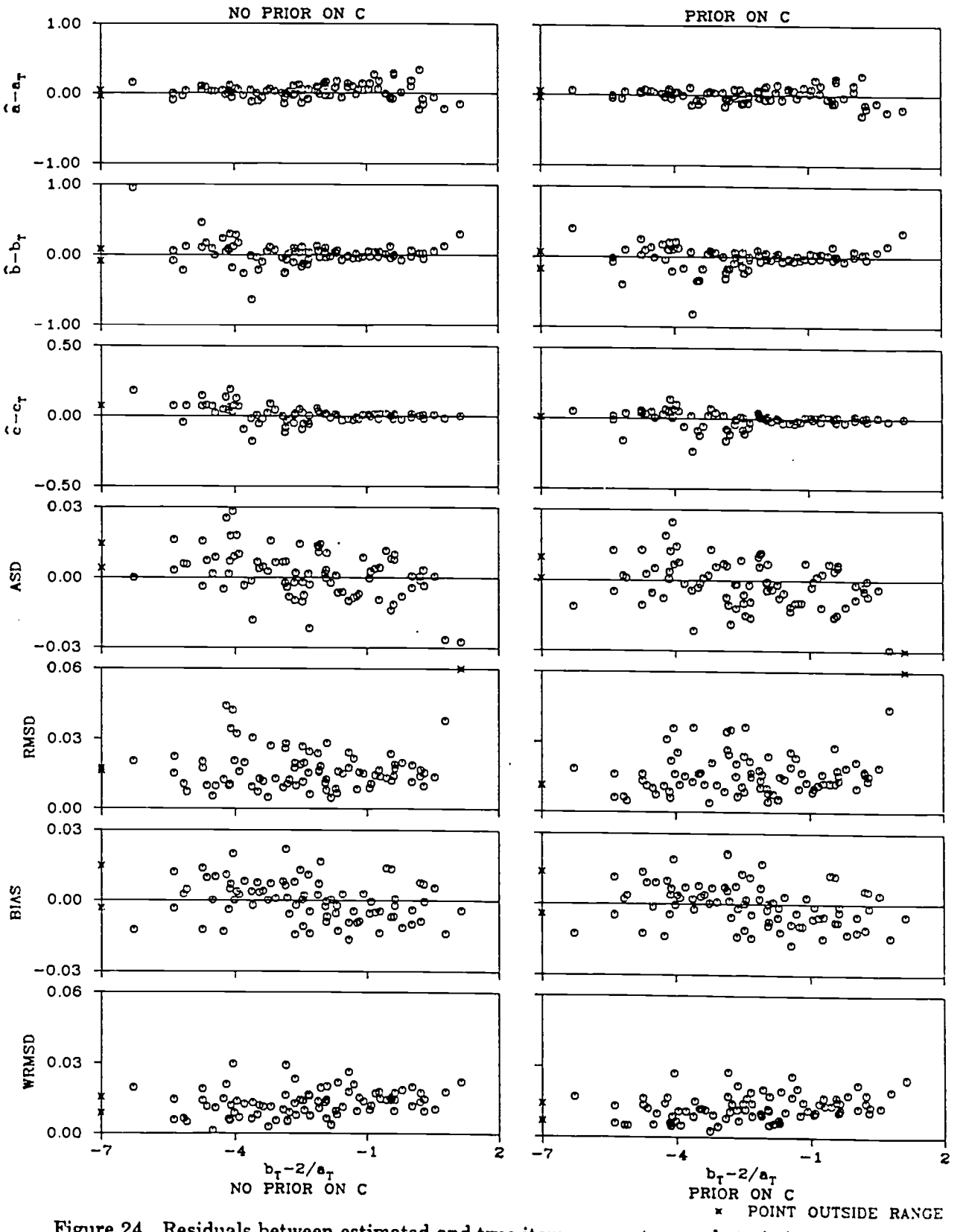

Figure 24. Residuals between estimated and true item parameters and statistics comparing estimated IRF's to true IRF's for the Method = with no prior on c and the Method P with a prior on c. All results plotted against the true b-2/a. Test R2.

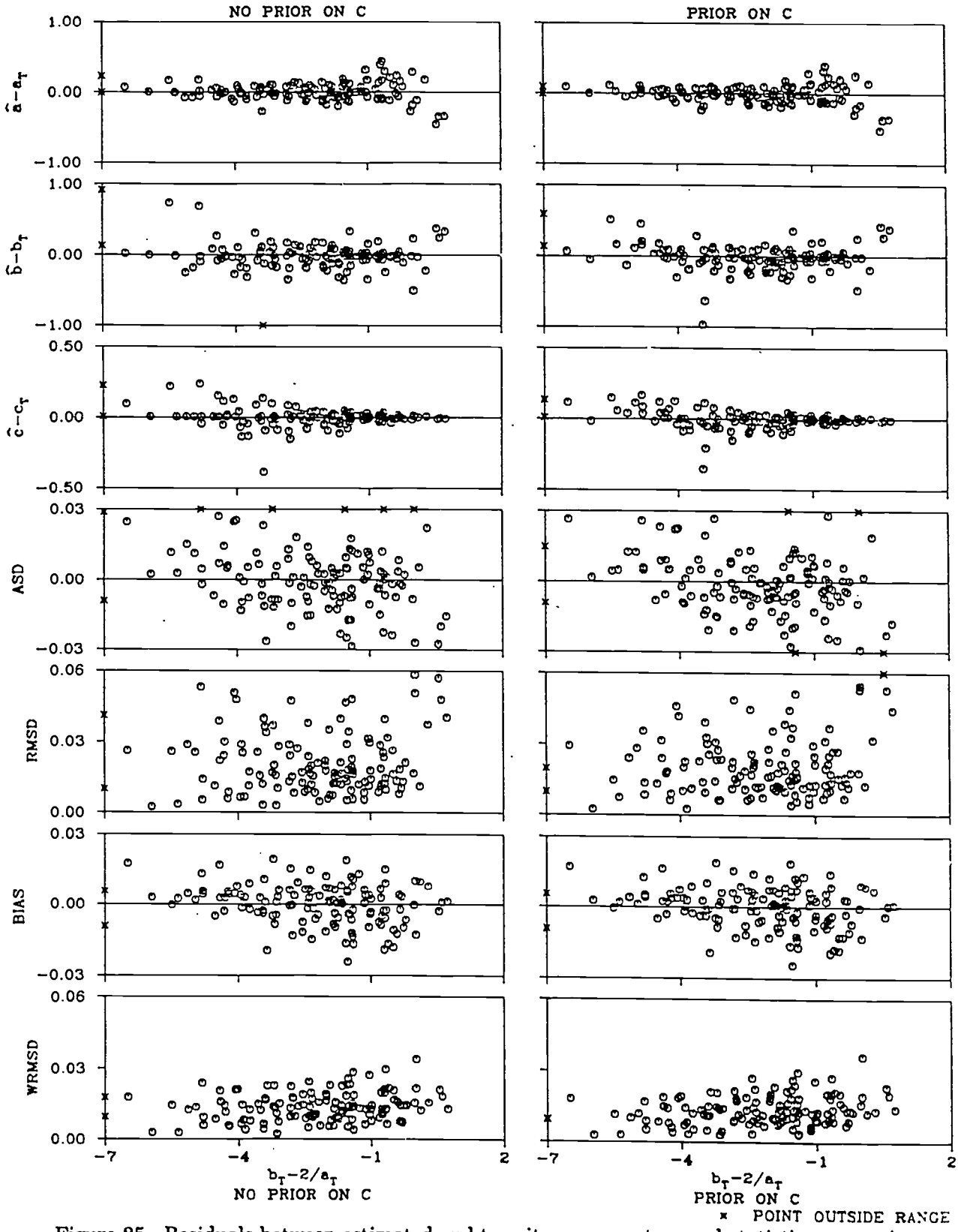

Figure 25. Residuals between estimated and true item parameters and statistics comparing estimated IRF's to true IRF's for the Method = with no prior on c and the Method P with a prior on c. All results plotted against the true b-2/a. Test R3.

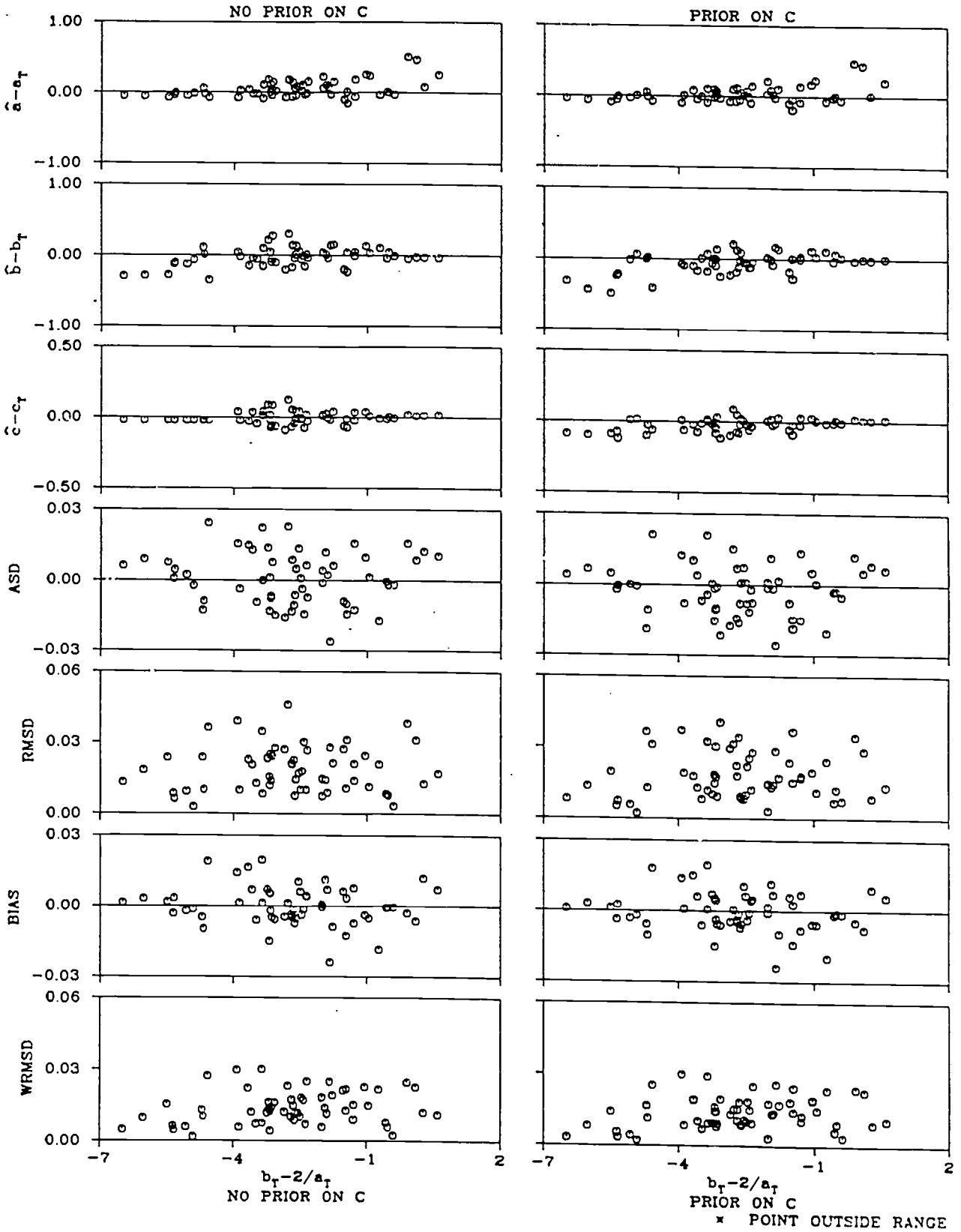

Figure 26. Residuals between estimated and true item parameters and statistics comparing estimated IRF's to true IRF's for the Method = with no prior on c and the Method P with a prior on c. All results plotted against the true b-2/a. Test R4.

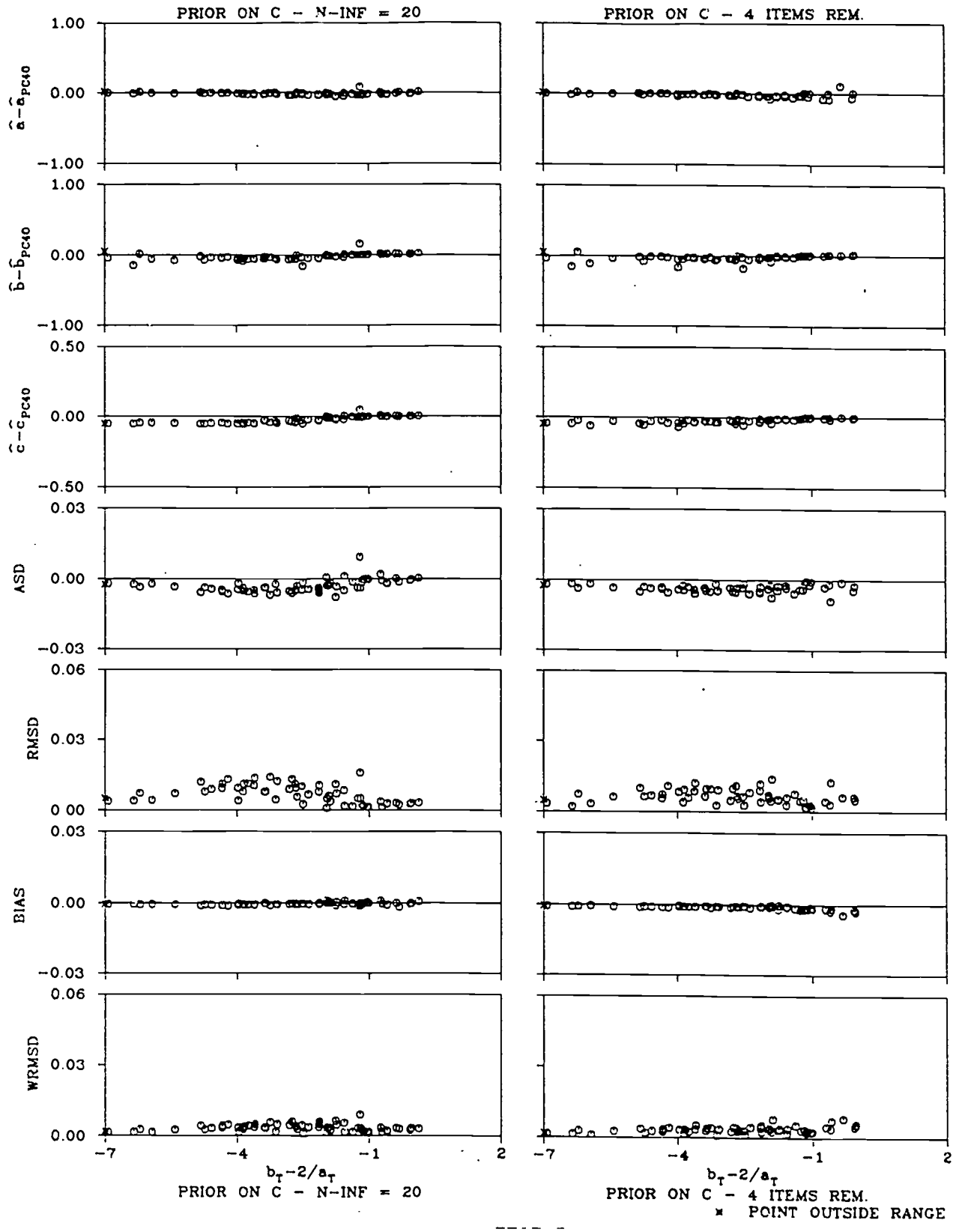

Figure 27. Test R1. Prior on c. Comparison of parameters and IRF's for calibration with $N^*_{-\infty}$ equal to 20 to calibration with $N^*_{-\infty}$ equal to 40 and for calibration with four items removed with $N^*_{-\infty}$ equal to 40 to calibration with all items with $N^*_{-\infty}$ equal to 40. Results plotted against true b-2/a.

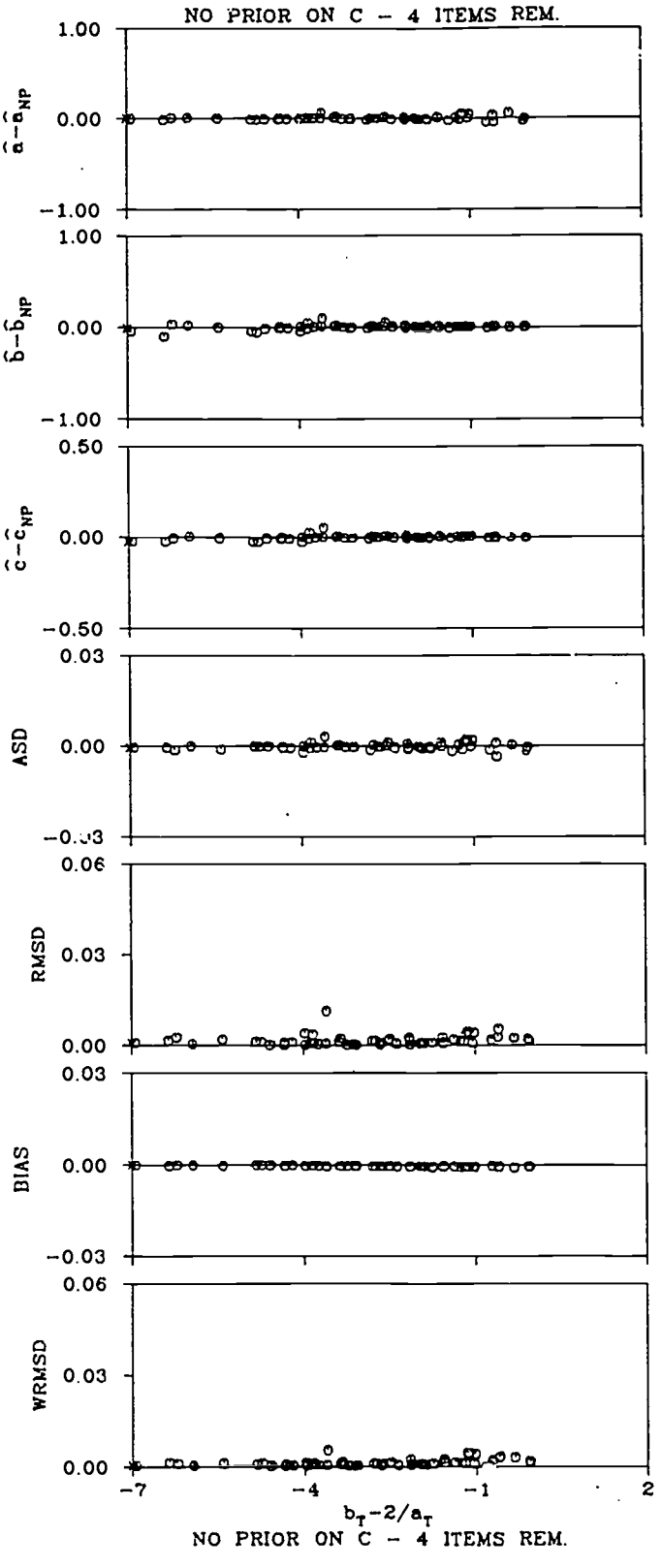

Figure 28. Test R1. No prior on c. Comparison of parameters estimated with all items and with four items removed.

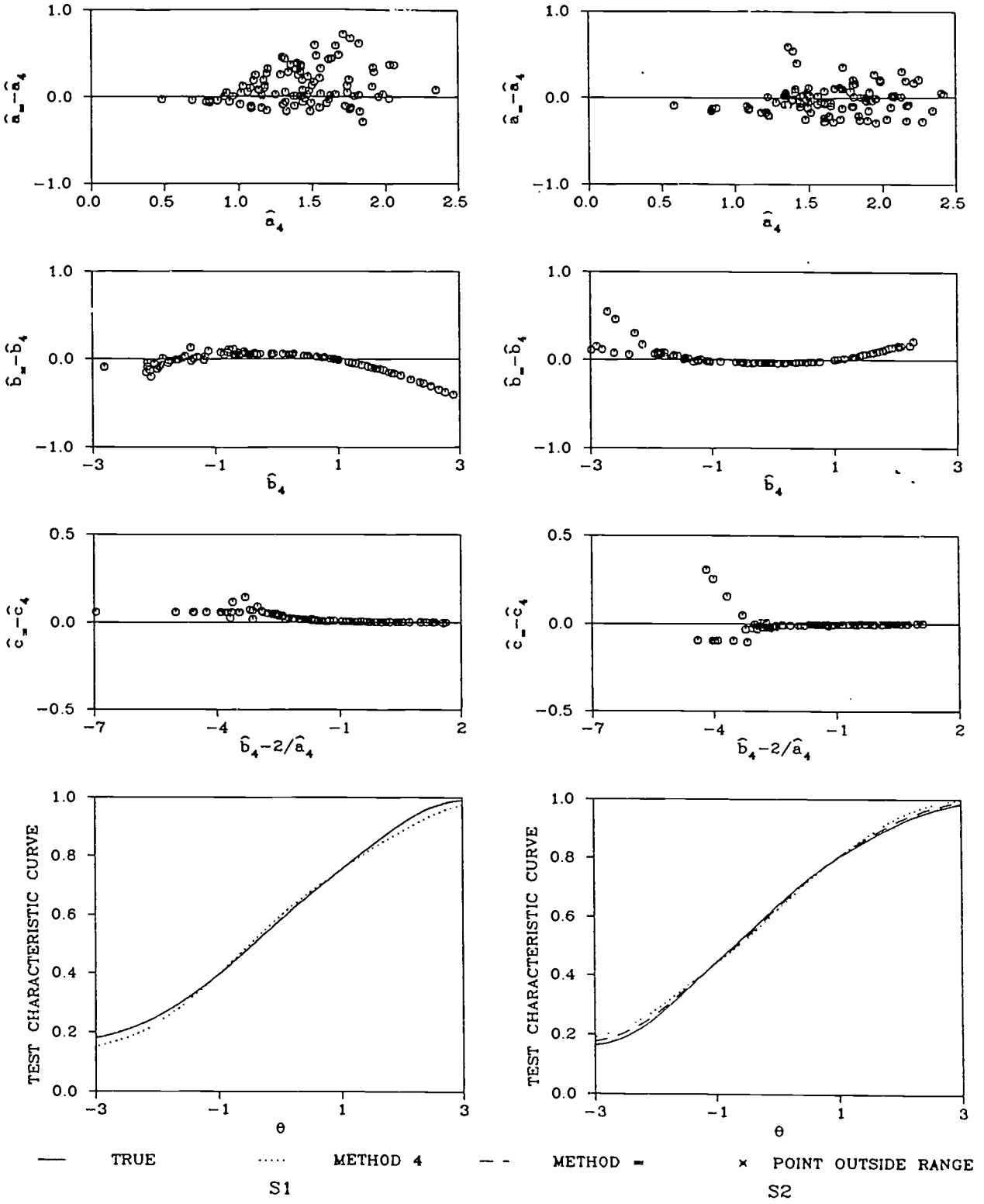

Figure 29. Residuals between the item parameter estimates for Method 4 and Method = and a plot of three test characteristic curves where the solid line is the true curve, the dotted line is Method 4 and the dashed line is Method = . Column 1 is for Test S1 and column 2 is for Test S2.

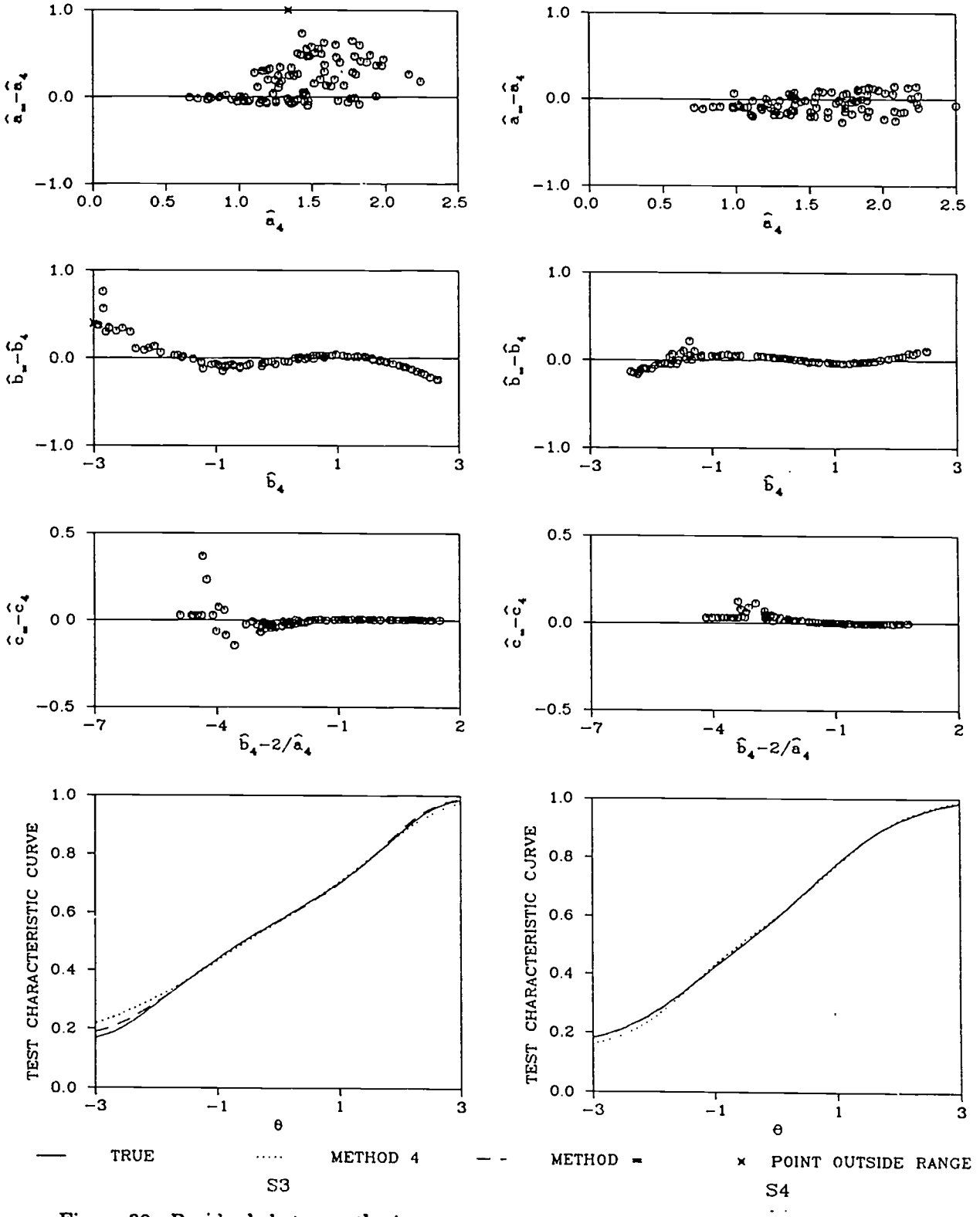

Figure 30. Residuals between the item parameter estimates for Method 4 and Method = and a plot of three test characteristic curves where the solid line is the true curve, the dotted line is Method 4 and the dashed line is Method = . Column 1 is for Test S3 and column 2 is for Test S4.

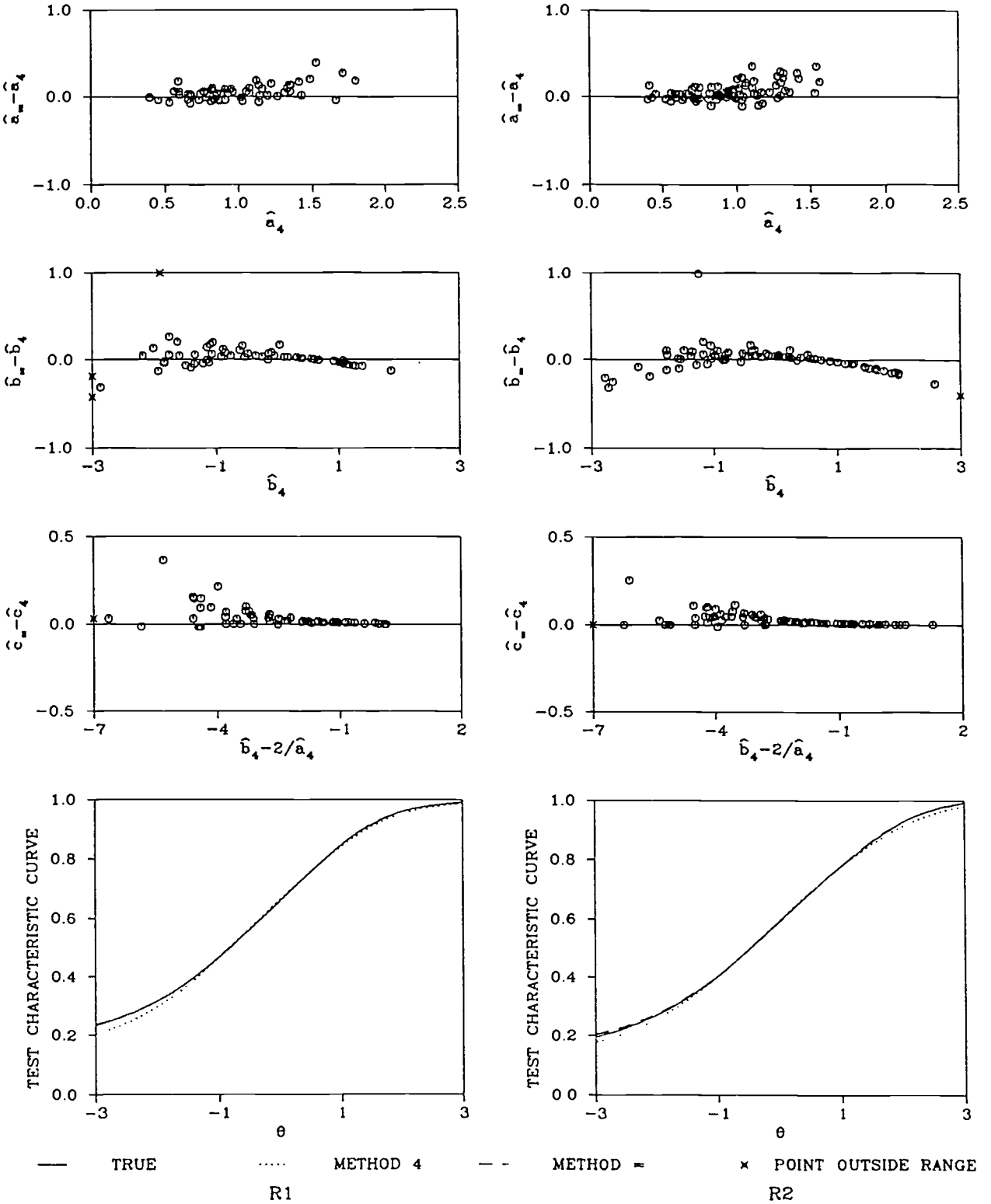

Figure 31. Residuals between the item parameter estimates for Method 4 and Method = and a plot of three test characteristic curves where the solid line is the true curve, the dotted line is Method 4 and the dashed line is Method = . Column 1 is for Test R1 and column 2 is for Test R2.

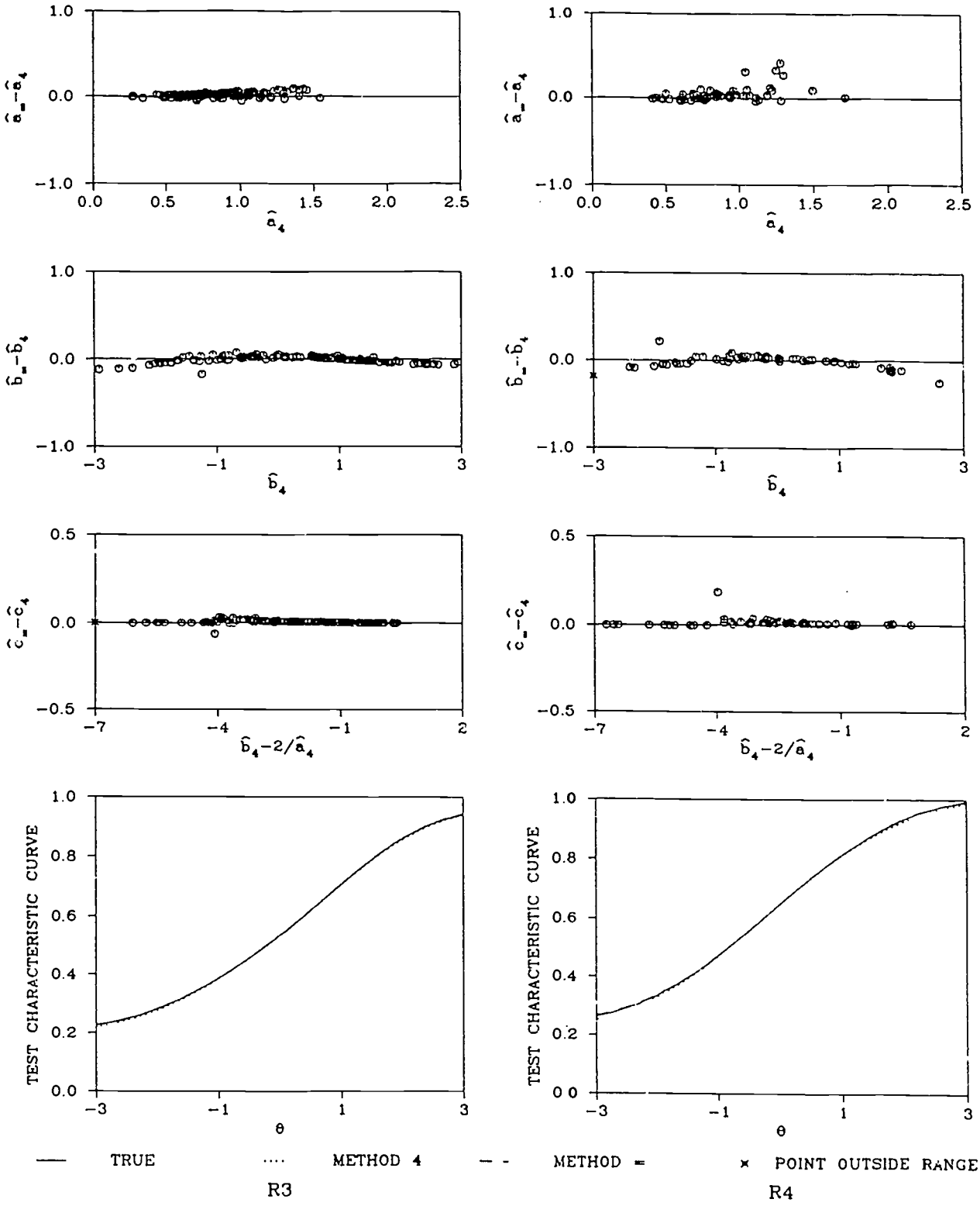

Figure 32. Residuals between the item parameter estimates for Method 4 and Method = and a plot of three test characteristic curves where the solid line is the true curve, the dotted line is Method 4 and the dashed line is Method = . Column 1 is for Test R3 and column 2 is for Test R4.

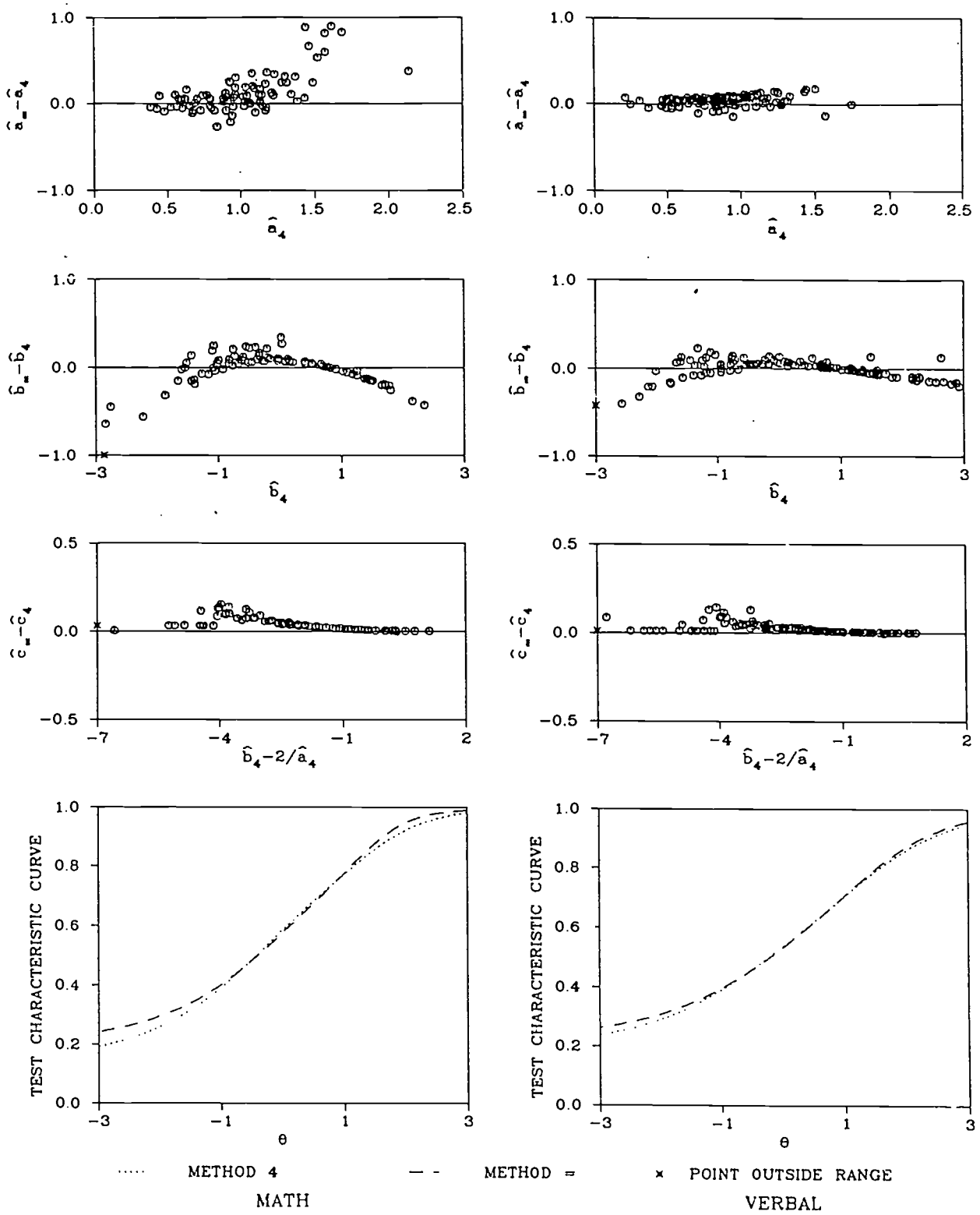

Figure 33. Residuals between the item parameter estimates for Method 4 and Method = and a plot of three test characteristic curves where the solid line is the true curve, the dotted line is Method 4 and the dashed line is Method = . Column 1 is for MATH and column 2 is for VERBAL.